ONLINE APPENDIX

Table of contents

Results expressed as Oaxaca-Blinder decompositions (expanded summary)

The Oaxaca-Blinder decomposition[1,2] is often used in labor economics to understand how much differences in group-characteristics explain disparities (or differences, more generally) in outcomes across groups e.g. disparities in log-wages across women vs. men, blacks vs. whites, union-members vs. non-members, etc. It partitions the total wage-difference into a portion due to differences in the distribution of potentially explanatory variables (termed the explained portion or composition effect), and a residual portion that cannot be explained by differences in these variables (termed the unexplained portion or structure effect). The unexplained portion is referred to as the structure effect because, as we will show, it captures the extent to which associations between the explanatory variables and the outcome vary across groups. Though the Oaxaca-Blinder decomposition was first introduced using linear models to decompose mean differences, more general forms have been introduced to decompose non-linear outcomes (Yun 2004[3]; Farlie 2005[4]), quantiles (Van Kerm 2015[5]), and the entire distribution of the outcome (Rothe 2015[6]) through various weighting (Dinardo 1996[7]; Barsky 2002[8]; Kline 2011[9]; Elder 2015[10]; Słóczynski 2015[11]), matching (Black 2006[12]), and other techniques (Firpo 2007[13]).

Here, we outline the Oaxaca-Blinder decomposition under linear models for the mean and discuss the conditions under which the Propositions 1 through 4 can be viewed as a causal version of the Oaxaca-Blinder decomposition with respect to interventions to set the distributions of the explanatory variables. The non-parametric results for propositions 1 through 4 could likewise be used to establish causal interpretations for the more general Oaxaca-Blinder decomposition methods for the mean,[8,12,14] with the results for propositions 5 through 7 extending them to the case of time-dependent confounding. Previous literature has for the most part concerned causal inference with respect to interventions to set group membership e.g. race, gender, union-membership etc[6,8,9,12,14,15] (see pages 11-15 for details). With the exception of certain study designs,[16,17] the interpretation of such an intervention is more difficult when the variable to be intervened upon is race or gender. We consider an alternative causal interpretation below.

A review of marginal Oaxaca-Blinder decompositions

Let us consider an Oaxaca-Blinder decomposition to estimate the portion of the racial disparity in log-wages Y that is statistically explained vs. not explained by racial differences in variables $V_1 \ldots V_n$, where the comparison across race concerns blacks R=1 vs. whites R=0. A typical Oaxaca-Blinder decomposition would proceed by fitting two race-specific regressions for the outcome given the explanatory variables:

$$E[Y|R=1,v] = \beta_0^{R=1} + \beta_1^{R=1}v_1 + \beta_2^{R=1}v_2 + \ldots + \beta_n^{R=1}v_n$$
$$E[Y|R=0,v] = \beta_0^{R=0} + \beta_1^{R=0}v_1 + \beta_2^{R=0}v_2 + \ldots + \beta_n^{R=0}v_n$$

Along with these we would estimate the mean value of each explanatory variable V_j among whites e.g. $E[V_1|R=0],\ldots,E[V_n|R=0]$. Then, the typical Oaxaca-Blinder decomposition expresses the marginal racial disparity in mean log-wages as a function of the explanatory variables' means (among whites) and their race-specific regression parameters:

$$E[Y|R=1] - E[Y|R=0]$$
$$= (\beta_0^{R=1} - \beta_0^{R=0}) + \sum_{j=1}^{n}(\beta_1^{R=1} - \beta_0^{R=0})E[V_j|R=0] + \sum_{j=1}^{n}\beta_j^{R=1}\{E[V_j|R=1] - E[V_j|R=0]\}$$

In what is called the aggregate decomposition,[14] the goal is to understand the extent to which the racial disparity is statistically explained by the fact that racial groups have different means for the

explanatory variables. The term $\sum_{j=1}^{n} \beta_j^{R=1}\{E[V_j|R=1] - E[V_j|R=0]\}$ comprises the 'explained portion' or what is also called the 'composition effect.' It captures racial differences in the mean values of the explanatory variables. The sum of the terms ($\beta_0^{R=1}$- $\beta_0^{R=0}$) and $\sum_{j=1}^{n}(\beta_j^{R=1} - \beta_j^{R=0})E[V_j|R=0]$ comprises the 'unexplained portion' or what is also called the 'structure effect.' It captures the portion of the disparity that cannot be statistically explained by differences in the means of explanatory variables i.e. differences in mean log-wages at the reference levels of the explanatory variables and also racial differences in the associations between each explanatory variable and the mean of the outcome log-wages.

In what is called the detailed decomposition,[14] the term $\beta_j^{R=1}\{E[V_j|R=1] - E[V_j|R=0]\}$ is interpreted as the independent contribution of the explanatory variable V_j to the 'explained portion' i.e. the portion of the disparity that is statistically attributable to the fact that the mean of V_j differs across racial groups (independently of racial differences in the means of the other explanatory variables). The term $(\beta_j^{R=1} - \beta_j^{R=0})E[V_j|R=0]$ is interpreted as the contribution of V_j to the 'unexplained portion' i.e. the portion of the disparity that is statistically explained by differences in the race-specific associations between the explanatory variable V_j and the mean of log-wages.

Defining conditional Oaxaca-Blinder decompositions

The typical Oaxaca-Blinder decomposition concerns the marginal racial disparity in log-wages E[Y|R=1]- E[Y|R=0], but one can extend it to decompose the racial disparity within levels of conditioning variables C i.e. E[Y|R=1,c]- E[Y|R=0,c]. These conditioning variables differ from explanatory variables V_j in that they are used to define the population rather than to explain the disparity. To accomplish this, one first fits race-specific models for the mean of log-wages given the explanatory variables V_j and also the conditioning variables C.

$$E[Y|R=1,v,c] = \beta_0^{R=1,c} + \beta_1^{R=1,c}v_1 + \beta_2^{R=1,c}v_2 + \ldots + \beta_n^{R=1,c}v_n + \beta_c^{R=1,c}c$$
$$E[Y|R=0,v,c] = \beta_0^{R=0,c} + \beta_1^{R=0,c}v_1 + \beta_2^{R=0,c}v_2 + \ldots + \beta_n^{R=0,c}v_n + \beta_c^{R=0,c}c$$

It can be shown that, the disparity within levels of C can be expressed as a function of the means of explanatory variables V_j given C=c and also the regression parameters that also condition on C=c:

$$E[Y|R=1,c]- E[Y|R=0,c]$$
$$= \left(\beta_0^{R=1,c} - \beta_0^{R=0,c}\right) + \sum_{j=1}^{n}(\beta_j^{R=1,c} - \beta_j^{R=0,c})E[V_j|R=0,c] + \sum(\beta_{c'}^{R=1,c} - \beta_{c'}^{R=0,c})c' +$$
$$\sum_{j=1}^{n}\beta_j^{R=1,c}\{E[V_j|R=1,c] - E[V_j|R=0,c]\}$$

In an aggregate decomposition,[14] we can consider the term $\sum_{j=1}^{n}\beta_j^{R=1,c}\{E[V_j|R=1,c] - E[V_j|R=0,c]\}$ to comprise the 'explained portion' because it captures the portion of the disparity that is statistically attributable to the fact that the means of explanatory variables differs across racial groups within levels of C. We can consider the terms $\left(\beta_0^{R=1,c} - \beta_0^{R=0,c}\right)$ and $\sum_{j=1}^{n}(\beta_j^{R=1,c} - \beta_j^{R=0,c})E[V_j|R=0,c]$ and $\sum(\beta_{c'}^{R=1,c} - \beta_{c'}^{R=0,c})c'$ to comprise the 'unexplained portion' because it captures the portion of the disparity that is statistically explained by the fact that associations between the explanatory variables and mean log-wages, and also the associations between the conditioning variables C and mean log-wages, differ by race. It follows then, that if there is no statistical association between race and the covariates C, the conditional decomposition has the same form as the marginal decomposition except that its components are specific to the levels of the conditioning variables C=c. We could also interpret the components for each explanatory

variable V_j in a detailed decomposition[14] as we did so in the marginal decomposition but again, these interpretations would pertain to a specific level of the conditioning variables C=c.

To the best of our knowledge, we have not seen such conditional forms of the Oaxaca-Blinder decomposition considered in the economics or epidemiology literature. While this extension is relatively minor, it has important implications when it comes to the causal interpretation of the decompositions. As described below if an explanatory variable V_j is exchangeable given the conditioning variables C and other explanatory variables that temporally precede V_j, then this permits causal inference where the 'explained' portion represents the disparity reduction under an intervention to equalize the explanatory variables V_j, and the 'unexplained' portion represents the corresponding residual disparity. The intervention does not completely eliminate disparities that arise due to any heterogeneity for the effect of V_j on Y across groups R, as this contributes to the residual disparity. We outline this for Propositions 1-4 below and provide all supporting proofs below. The causal interpretations given here are thus with respect to the explanatory variables V, rather than to hypothetical interventions on race itself as per other literature.[8,9,12,14,15]

Propositions 1-4 expressed as causal implementations of the Oaxaca-Blinder decomposition

Suppose now that we fit three sets of regressions:

Set 1:
$E[Y|R=1,x,c]=\omega_0 + \omega_1 x + \omega_3'c$
$E[Y|R=0,x,c]=\pi_0 + \pi_1 x + \pi_3'c$

Set 2:
$E[Y|R=1,m,x,c]=\alpha_0 + \alpha_1 x + \alpha_2 m + \alpha_3'c$
$E[Y|R=0,m,x,c]=\beta_0 + \beta_1 x + \beta_2 m + \beta_3'c$

Set 3:
$E[Y|r,x,m,c] = \theta_0 + \theta_1 r + \theta_2 x + \theta_3 m + \theta_4 rx + \theta_5 rm + \theta_6'c$
$E[Y|r,x,c] = \gamma_0 + \gamma_1 r + \gamma_2 x + \gamma_4 rx + \gamma_6'c$
$E[Y|r,c] = \phi_0 + \phi_1 r + \phi_6'c$

Suppose further that, for simplicity but not out of necessity, we assume no statistical interactions between race R and covariates C for the mean outcome log-wages i.e. $\omega_3 = \pi_3$ and $\alpha_3 = \beta_3$, such that the models of set 1 are equivalent to the second model in set 3, and the models in set 2 are equivalent to the first model in set 3. The models are equivalent in the sense that they allow for heterogeneous effects of childhood SES X and test scores M across race R. Note that all of our arguments assume no interactions between race and conditioning covariates C, but this is only done to simplify the proofs below.

Goal of Proposition 1: equalize childhood SES across race given covariates i.e. standardization

We can carry out an aggregate Oaxaca-Blinder decomposition to understand the extent to which differences in childhood SES X statistically explain the racial disparity within levels of gender and age C. With the models in set 1, the unexplained portion equals $(\omega_0 - \pi_0) + (\omega_1 - \pi_1)E[X|R=0,c]$, and the explained portion equals $\omega_1\{E[X|R=1,c] - E[X|R=0,c]\}$. Now, consider the linear models of set 3 and assume that the effect of childhood SES X on log-wages is unconfounded given covariates gender and age C=c holds (assumption A1). Under Proposition 1, an intervention to set the distribution of childhood SES X among blacks according to its distribution among whites with

covariates C=c, we have that the residual disparity equals: $\gamma_1 + \gamma_4$ E[X|R=0,c], and the disparity reduction equals: $(\gamma_2 + \gamma_4)$ {E[X|R=1,c] - E[X|R=0,c]}. We show in the proofs below that the unexplained portion and the residual disparity are equal, and likewise the explained portion and the disparity reduced are equal.

Goal of Proposition 2: equalize test scores across race given childhood SES and covariates i.e. mediation-analysis

We can carry out an aggregate Oaxaca-Blinder decomposition to understand the extent to which differences in test scores M statistically explain the racial disparity within levels of childhood SES X, gender and age C. With the models in set 2, the unexplained portion equals: $(\alpha_0-\beta_0) + (\alpha_1-\beta_1)x + (\alpha_2-\beta_2)$E[M|R=0,x,c], and the explained portion equals: α_2{E[M|R=1,x,c] - E[M|R=0,x,c]}. Now, consider the linear models of set 3 and assume that the effect of test scores M on log-wages is unconfounded given childhood SES X and covariates gender and age C=c holds (i.e. assumption A2). Under Proposition 2, an intervention to set the distribution of test scores M among blacks according to its distribution among whites with childhood SES X=x and covariates gender and age C=c, we have that the residual disparity is equal to $\theta_1 + \theta_4 x + \theta_5$E[M|R=0,x,c], and the disparity reduction is equal to $(\theta_3 + \theta_5)${E[M|R=1,x,c] - E[M|R=0,x,c]}. We show in the proofs below that the unexplained portion and the residual disparity are equal, and likewise the explained portion and the disparity reduced are equal.

Goal of Proposition 3: equalize childhood SES and test scores across race given covariates

We can carry out an aggregate Oaxaca-Blinder decomposition to understand the extent to which differences in childhood SES X and test scores M statistically explain the racial disparity within levels of covariates gender and age C. With the models in set 2, the unexplained portion equals: $(\alpha_0-\beta_0) + (\alpha_1-\beta_1)$E[X|R=0,c] + $(\alpha_2-\beta_2)$E[M|R=0,c], and the explained portion equals: α_1{E[X|R=1,c] - E[X|R=0,c]} + α_2{E[M|R=1,c] - E[M|R=0,c]}. Now, consider the linear models of set 3 and assumptions A1 and A2. Under Proposition 3, an intervention to set the distribution of childhood SES X and test scores M among blacks according to their distribution among whites with covariates C=c, we have that the residual disparity is equal to $\theta_1 + \theta_4$E[X|R=0,c] + θ_5E[M|R=0,c], and the disparity reduction is equal to $(\theta_2 + \theta_4)$ {E[X|R=1,c] - E[X|R=0,c]} + $(\theta_3 + \theta_5)$ {E[M|R=1,c] - E[M|R=0,c]}. We show in the proofs below that the unexplained portion and the residual disparity are equal, and likewise the explained portion and the disparity reduced are equal.

Goal of Proposition 4: equalize test scores across race given covariates

We can carry out a detailed Oaxaca-Blinder decomposition to understand the extent to which differences in childhood SES X, and also differences in test scores M, each statistically explain, independent of each other, the racial disparity within levels of gender and age C. With the models in set 2, the part of the unexplained portion due to racial differences in the association between childhood SES X and log-wages equals $(\alpha_1-\beta_1)$E[X|R=0,c]; the part of the unexplained portion due to racial differences in the association between test scores M and log-wages equals $(\alpha_2-\beta_2)$E[M|R=0,c]; the part of the explained portion due to racial differences in the distribution of childhood SES X (independent of racial differences in test scores M) equals α_1{E[X|R=1,c] - E[X|R=0,c]}; the part of the explained portion due to racial differences in the distribution of test scores M (independent of racial differences in childhood SES X) equals α_2{E[M|R=1,c] - E[M|R=0,c]}. Now, consider the linear models of set 3 and assumption A2. Under Proposition 4, an intervention to set the distribution of test scores M among blacks according to their distribution among whites with covariates gender

and age C=c, we have that the residual disparity is equal to $\theta_1 + \theta_2\{E[X|R=1,c]-E[X|R=0,c]\} + \theta_4 E[X|R=1,c]+ \theta_5 E[M|R=0,c]$, and the disparity reduction is equal to $(\theta_3 + \theta_5)\{E[M|R=1,c] - E[M|R=0,c]\}$. We show in the proofs below that the portion explained independently by test scores M and the disparity reduction are equal, and that the sum of the entire unexplained portion and the portion independently explain by childhood SES X equals the disparity reduced.

A further note about causal interpretation under the detailed decomposition

Note that the detailed decomposition interprets $\alpha_1\{E[X|R=1,c] - E[X|R=0,c]\}$ as the portion of the disparity in log-wages statistically explained by racial differences in the mean of childhood SES X given covariates C gender and age (independent of racial differences in test scores M). However, this does not in general equal the disparity reduction under Proposition 1 i.e. what would occur under equalizing the distribution of childhood SES X across race given covariates C. Only when the effect of childhood SES X on log-wages is not mediated by test scores M, such that $\omega_1=\alpha_1$, would this interpretation apply. Otherwise it is not clear what the causal interpretation is for a detailed decomposition regarding childhood SES X in the models from set 2. Thus, the causal interpretation of a detailed decomposition depends on the causal structure among the explanatory variables V_j, which is related to the 'path-dependence' issue long noted by economists (Fortin 2011[14]; Rothe 2015[6]). Here we suggest an approach for when separate causal interpretations for each explanatory variable in the Oaxaca-Blinder decomposition are desired. Considering the hypothesized causal structure carefully, one can carry out proposition 4 separately for each V_j of interest, including only additional variables that precede V_j that suffice to control for confounding of V_j. These set of results would answer how the disparity would change under alternative interventions to equalize each V_j marginally. See below 'Relations to the literature on causal inference and the Oaxaca-Blinder decomposition' for further considerations.

A further note about the choice of referent group

In our intervention of interest, X and/or M are assigned among blacks R=0 according to the distribution found among whites R=0 (the referent group). We could have alternatively chosen R=1 as the referent group i.e., an intervention to assign X and/or M among whites R=0 according to the distribution found among blacks R=1. We also could have chosen to assign X and/or M to each R=1 and R=0 according to the distribution found among the combined population of R=1 and R=0. Following the logic of our proofs, it can be shown that each choice can yield an Oaxaca-Blinder type decomposition. When outcomes for blacks fare worse than whites, the choice of R=0 as the referent would constitute a disparity reduction through improvement for blacks rather than declines for whites. We suspect that, of these considerations, this proposal is likely of most interest to policy-makers in the United States examining solutions to eliminate racial disparities in the United States.

Overview of relations to the broader decomposition methods literature

In this section, we place our contribution in the context of decomposition methods from the causal mediation literature from epidemiology and biostatistics, and the Oaxaca-Blinder decomposition literature from economics. Although our motivating example concerned decompositions of racial/ethnic differences, our methods could be adapted to examine differences across other social groupings e.g. across gender, sexual orientation, socioeconomic classes in childhood, or even groups that are defined later in life e.g. socioeconomic class in adulthood, health insurance status or type, or union group membership. Our manuscript thus presents a causal decomposition for any group-outcome association. It does not pose counterfactuals for setting group status, which will often be non-manipulable as in the case of race/ethnicity (Holland 1986[18]), but rather for setting a manipulable target variable that may occur concurrently with group status, or may be an effect of group status. Our results can appropriately account for variables affected by group status that confound the effect of the target on the outcome. For the remainder of our discussion we will focus on racial classification as the group defining variable. Herein, also, we focus our discussion on the assumption of conditional exchangeability (Hernan & Robins 2018[19]) that was at the foundation of our contribution. Other assumptions such as consistency and positivity are also critically important to consider.[19]

Implicit causal model

For pedagogical purposes, let us for clarity define a causal graph in the sense of Robins 2011,[20] similar to the one provided in Figure 1. On this graph, we define the variables M as test scores, X as childhood SES, R as race with r=black vs. r^*=white, H as historical structures of racism, C as covariates gender and age. Now, on this graph, Y is a direct descendant of R, X, M, C; M is a direct descendant of R, X, C; X is a direct descendant of H, C; and R is direct descendant of H. We will sometimes make reference to the possible presence of a variable L that when envoked is a direct descendant of R, X and C as well as a direct parent of M and Y. We also allow for a selection node denoting membership in the sample at baseline that is a direct descendant of R, C, X, H. This selection node S is conditioned upon, representing collider-stratification[21] induced associations between race, gender and age that arise (in the sample of non-institutionalized at baseline) through selective forces e.g. early life mortality and incarceration.

There are at least two causal models that could be applied to provide interpretations for this graph. Suppose there are K variables V, and that we index their temporal order with subscripts such that V_{k-1} always precedes V_k. Let PA_k denote the direct parents of V_k. We use an overbar $\bar{V}_k$ to denote the vector of values of V through node k $(V_1, \dots, V_k)$. We define the counterfactual $V_k(v_{k-1})$ as the value that variable V_k would obtain under an intervention to set V_{k-1} to value v_{k-1}. The Minimal Causal Model (MCM) of Robins 2011[20] posits that the "one-step ahead" counterfactuals $V_k(v_{k-1})$ exist; any variable V_k is only a function of its past through the observed values of its direct parents i.e. $V_k(\bar{V}_{k-1}) \equiv V_k(pa_k)$; the counterfactuals are recursively obtained; given the factual past of V_k i.e., $\bar{V}_{k-1} = \bar{v}_{k-1}$, the distribution of the one step ahead counterfactuals $V_{k+1}(v_k)$ and their propagation are independent of $V_k = v_k$. The Non-Parametric Structural Equation Model (NPSEM) of Pearl 2009[21] realizes each variable as an unspecified, deterministic function of its parents' values and an error term: $Y = f(r, m, x, c, e_y)$, $M = f(r, x, c, e_m)$, $X = f(h, c, e_x)$, $R = f(h, e_r)$, $S = f(r, c, x, h, e_s)$, wherein the error terms e_v are mutually independent. H is left unspecified. With L present, we have $Y = f(r, m, l, x, c, e_y)$, $M = f(r, l, x, c, e_m)$, $L = f(r, x, c, e_l)$, $X = f(h, c, e_h)$, $R =$

$f(h, e_r)$, $S = f(r, c, x, h, e_s)$. In this model counterfactuals are also recursively obtained, and any variable V_k is independent of its past given the values (factual or otherwise) of its parents PA_k. Under both models the distribution of V can be written as $f_v(V) = \prod_{j=1}^{K} f(v_j|pa_j)$. A key distinction is that the MCM is agnostic about the existence of cross-world counterfactuals e.g., $Y(X = 1, M(X = 0))$ whereas the NPSEM explicitly permits them. Our proposals are compatible with either interpretation of the causal graph. Our formal results, however, do not reference any particular graph. Thus, we provide this graph for intuition only.

Let us consider now the substantive assumptions implied by the absence of arrows on this graph. For example, the graph assumes that race R does not affect childhood SES. In our conceptualization, racial classification could be considered to be determined at the time of conception as a function of the parents' own racial classification and societal norms at the time of the survey. Racial classification is not a completely deterministic function of maternal or paternal race when we consider that children of mixed race were reported as such in the NLSY97 survey. Concerning SES during childhood, we assume that it would have been derived from maternal SES at that time. As a function of maternal SES, therefore, SES during childhood would not result from discrimination during childhood, before entry into the labor market or establishment of an occupation. Any wealth held during childhood is likely to be inherited and unlikely to be used for securing material resources for wellbeing during childhood. Moreover, both racial classification and parental SES during childhood are determined by the historical structuring of society's wealth and opportunities enacted through slavery, Jim Crow, federal housing policy, and other structural forms of racism and discrimination.[22–24]. Thus, parental features are subsumed into the historical processes node. These arguments strongly weigh against a direct effect of race on childhood SES but do allow for an association between race and SES through historical processes. Historical processes are assumed to completely exert their effects on the outcome through race and socioeconomic status in early life i.e., no direct effect of historical processes with respect to race, childhood SES and test scores. This seems to be a reasonable assumption as SES is a primary source of social stratification in the U.S. in addition to race and gender. Also absent from the graph are any unmeasured common determinants of test scores M and the outcome Y. This assumption is more tenuous as there may be heterogeneity in school quality and family context even within the same socioeconomic bracket, age, gender, race, and ethnicity group and this heterogeneity may be predictive of the outcome.

Relations to the causal mediation analysis literature

The natural direct effect is defined as $E[Y(r, M(r^*))|c] - E[Y(r^*, M(r^*))|c]$ and the natural indirect effect is defined as $E[Y(r, M(r^*))|c] - E[Y(r, M(r))|c]$, and marginal effects can be obtained by standardizing each mean counterfactual to the same covariate distribution e.g., $P[C = c]$.[21,25,26] The counterfactual $Y(r, M(r))$ represents the outcome that would be observed under an intervention to jointly set an individual's (i) race to value r (ii) mediator to value $M(r)$ it would obtain under an intervention to set that person's race to value r. Holland 1986[27] argues that such counterfactuals are difficult to imagine and estimate for non-manipulable characteristics such as race. Holland's concerns are amplified for natural direct/indirect effects because they rely on nested counterfactuals $Y(r, M(r^*))$ which in this context seem difficult to ascertain. This counterfactual represents the outcome for a person who was assigned to be racially classified as black, but having

the test score that would have been observed for this person under an alternative life being classified as white.

If, in spite of these challenges, one still wished to study causal mediation of the effects of assigning racial classification, one would need to the following conditions to hold: no race-outcome confounding $Y(r,m) \coprod R|C$, no test score-outcome confounding $Y(r,m) \coprod M|R,C$, no race-test score confounding $M(r) \coprod R|C$, and no time-dependent confounding i.e. no common cause L of the test score-outcome relationship affected by race $Y(r,m) \coprod M(r^*)|C$. This last assumption is rather restrictive as it disallows the existence of other paths in which another mediator is a cause of the mediator of interest.[28] If there were such a confounder L of the $M-Y$ relationship such that it was affected by race, one would not be able to identify the natural direct/indirect effects even if L were measured.

To overcome the identification challenges posed by multiple mediators wherein one causes the other, methods have emerged that estimate natural direct/indirect effects with respect to L and test scores M jointly, under stricter no-confounding assumptions that apply to the set of L and M jointly: $Y(r,l,m) \coprod R|C$; $Y(r,l,m) \coprod L,M|R,C$; $L(r), M(r) \coprod R|C$; $Y(r,l,m) \coprod L(r^*), M(r^*)|C$. Methods have also been introduced for path-specific effects, including, in this context, what would be the direct effect of race on Y not through L or M ($i.e., R \rightarrow Y$), the effect of race through L (i.e., the sum of paths $R \rightarrow L \rightarrow M \rightarrow Y$ and $R \rightarrow L \rightarrow Y$), and also the effect of race on Y through its direct effect on M (i.e., $R \rightarrow M \rightarrow Y$).[28–32] Several sets of assumptions have been proposed to non-parametrically identify these effects but they are much stronger than those necessary for mediation with a single or joint set of mediators. Moreover, the effect of R on Y through M (i.e., the sum of paths $R \rightarrow M \rightarrow Y$ and $R \rightarrow L \rightarrow M \rightarrow Y$) are not identified without other restrictive assumptions such as the absence of unit-level interactions between R and L,M[33] or the appropriateness of linear models for L, M, and Y.[34,35] As with the methods for natural effects for a single mediator, these all concern counterfactuals for non-manipulable exposures. Some of the nested counterfactuals that define these effects are tremendously difficult to imagine for non-manipulable exposures e.g. $E[Y(r, L(r^*), M(r, L(r^*))|c]$ as with a path-specific effect $R \rightarrow M \rightarrow Y$.

Alternative methods for mediation analysis, or rather effect decomposition, are based on so-called randomized interventional analogues of natural direct effects $E\left[Y\left(r, G_{M(r^*)|c}\right)\right] - E\left[Y\left(r^*, G_{M(r^*)|c}\right)\right]$ and natural indirect effects $E\left[Y\left(r, G_{M(r)|c}\right)\right] - E\left[Y\left(r, G_{M(r^*)|c}\right)\right]$ where $G_{M(r^*)|c}$ represents the value of the mediator drawn from its distribution among the population with $C = c$ after assigning R to value r^*.[31] These analogues were originally developed in a non-counterfactual framework.[36,37] Later, a counterfactual framework was used to extend these analogues to the case where there exists another mediator L that serves to confound M's effect on Y,[31] where they were shown to be identified under weaker assumptions than natural direct effects: $Y(r,m) \coprod R|C$, $M(r) \coprod R|C$, and $Y(r,m) \coprod M|R,C,L$. In this context, the intervention is again drawn from among those with R set to r^* given $C = c$, irrespective of L. Alternate randomized interventional analogues have been proposed in contexts analogous to time-dependent confounding by L where the intervention depends on the value of $L(r^*)$ given $C = c$ after assigning race i.e. $G_{M(r^*)|r^*,L(r^*),c}$.[38] These variants of randomized interventional analogues have been extended to the case of path-specific effects under time-varying exposures and time-varying mediators[39,40] and path-specific effects under multiple mediators.[41–43] Each of these methods and their identification assumptions, however, still rely on assumptions necessary to identify the effect of assigning racial classification, while also assigning M according to its counterfactual distribution given C upon assigning race.

In the context where one is interested in decomposing a disparity, rather than the effect of racial classification, there is no need to satisfy the stringent identification assumptions posed by the aforementioned natural direct/indirect effects or their randomized interventional analogues. VanderWeele & Robinson 2014[44] essentially showed that progress can be made by defining a randomized intervention where the mediator M (or a baseline variable X) is drawn from its observed distribution among the referent racial group given C, and posing counterfactuals $Y(m)$ that only involve assignments for the mediator and not race, e.g., $G_{m|r^*,x,c}$. This substantially weakens the identification assumptions e.g. $E[Y(m)|m,r,c] = E[Y(m)|r,c]$ in the case of a target mediating variable M (similar assumptions are required for analogues based on a target baseline variable X). The identification results in VanderWeele & Robinson 2014[44] presume that there is no confounder L of the $M-Y$ relationship affected by race. These results were referenced in VanderWeele & Tchetgen Tchetgen 2016[39] to motivate the primary interest of randomized interventional analogues for time-varying exposures and mediators in the disparity setting. However, they only developed identification results under an assignment for the exposure (e.g. race in this context) under time-dependent confounding by L. They did not go on to develop identification results for the disparity setting wherein no assignment is envisioned for race.

A close inspection of estimators for the methods mentioned above will show that they equalize confounders across race R by conditioning upon baseline variables X, C or standardizing across X, C such that the disparity estimand pertains to a disparity where racial groups have either the same values for X, C e.g., $E[Y|R=1,x,c]-E[Y|R=0,x,c]$ or have the same distribution of X, C i.e., $\sum_{x,c} E[Y|R=1,x,c]P[X=x,C=c]-\sum_{x,c} E[Y|R=0|x,c]P[X=x,C=c]$. These methods ignore the fact that such measures of disparity where the effects of X are removed may not be meaningful or of substantive interest. Even though investigators may not wish to explain such disparities, the use of estimators based on the aforementioned methods will force their hand because such methods were designed to use X to resolve both race-outcome confounding and mediator-outcome confounding. This issue will become even more pervasive when considering disparities across statuses defined later in life such as health insurance status/type or union membership or geography.

Our decomposition in proposition 4 allows investigators to decompose a disparity that only conditions on a subset of baseline variables C, while controlling for mediator-outcome confounding by all baseline variables X, C. It accomplishes this by posing a randomized intervention $G_{m|r^*,c}$ to set M marginally with respect to the variable X but conditionally with respect to C. The disparity reduction and residual disparity under such an intervention are identified under milder assumptions (e.g., $E[Y(m)|m,x,r,c] = E[Y(m)|x,r,c]$) than those needed to identify the path-specific race effects described earlier. Proposition 7 extends this decomposition to the case where there is a confounder L of the $M-Y$ relationship that is affected by race; here the randomized intervention $G_{m|r^*,c}$ to set M is marginal with respect to L and X, under the assumption that $E[Y(m)|m,l,x,r,c] = E[Y(m)|l,x,r,c]$. Our decomposition in proposition 3 allows investigators to decompose a disparity that conditions on baseline variables C under a joint randomized intervention on a baseline variable X and a mediating variable M, assuming $E[Y(x,m)|m,x,r,c] = E[Y(x,m)|x,r,c]$ and $E[Y(x,m)|r,c] = E[Y(x,m)|x,r,c]$. Proposition 6 extends this decomposition to the case where both X and M are intervened upon and there is a confounder L of the $M-Y$ relationship that is either affected by R or X, under an intervention that is marginal with respect to L. Identification requires that assuming the effects of X and M are unconfounded i.e., $E[Y(x,m)|m,l,x,r,c] = E[Y(x,m)|l,x,r,c]$ and $E[Y(x,m)|r,c] = E[Y(x,m)|x,r,c]$.

We also generalize the randomized intervention on M in VanderWeele & Robinson 2014[44] that originally was used to decompose a disparity conditioned by X, C in the absence of time-dependent confounding. In our Proposition 5 above, we provide identification results for such a conditional disparity in the presence of a time-dependent confounder L of the $M - Y$ relationship that is affected by race. The non-parametric formulae for this case is equivalent to the formulae provided for certain randomized interventional analogues in the setting of time-dependent confounding by L under an assignment for racial classification .[31,39,43] However, in this context our estimand is identified under much weaker assumptions than those used to identify the decomposition of the total effect of race. Note that none of non-parametric formulae for our decompositions presented here pertain to those proposed by Zheng & Van Der Laan 2017.[40] In the context of our structural model above with time-dependent confounding, those quantities pose an intervention on race, drawing $M(r^*)$ from a distribution conditional on the counterfactual variable $L(r^*)$. If conditional interventions along these lines were of interest, an alternative approach would be to consider the disparity residual/reduction under a randomized intervention on M given observed L and all variables in X, C, a randomized interventional analogue identified in Jackson 2018.[45]

Relations to the literature on causal inference and the Oaxaca-Blinder decomposition

In the literature we review here, much of the work concerning Oaxaca-Blinder decompositions[6,8,9,11,12,14] pose a different causal model than the one we defined above. It typically poses unspecified non-parametric structural equations (generating functions) only for the outcome among each racial group $Y^r = f^r\left(m^r, x^r, c^r, e_y^r\right)$ and $Y^{r*} = f^{r*}\left(m^{r*}, x^{r*}, c^{r*}, e_y^{r*}\right)$. Given the explanatory variables $(\cdot)$ a counterfactual is posed by replacing $f^r\left(\cdot, e_y^r\right)$ with $f^{r*}\left(\cdot, e_y^{r*}\right)$ under the assumption that $e_y \coprod R|\,(\cdot)$.[14] This is interpreted as what would happen to a person with $R = r$ had they had the generating function of a person in group $R = r^*$. This counterfactual is identified on average if every structural determinant of Y that is associated with R has been included in the decomposition, and $E\left[e_y \middle| R, (\cdot)\right] = 0$. This assumption implies conditional exchangeability for race given the explanatory variables, i.e., $Y(r) \coprod R\,|(\cdot)$ for r, r^*. Thus, assuming a counterfactual generating function given a (hypothetically) sufficient set of explanatory variables is akin to hypothetically intervening to set R given fixed values for those explanatory variables. If in fact the explanatory variables contain all structural determinants of the outcome associated with race—including descendants of race and other determinants associated with these descendants—then this assumption would hold.[15] This is not the same counterfactual proposition for assigning race as is used with the natural direct/indirect effects,[15] yet it is still difficult to imagine for health outcomes (as opposed to healthcare services or treatment) as their determinants are often not fully known or measured. This counterfactual proposition does seem plausible for certain experimental designs.[16,17]

Aggregate Decomposition

For the aggregate decomposition, in our context, the goal is to apportion an observed group difference $E[Y|R = r] - E[Y|R = r^*]$ into that attributable to differences in the distribution of explanatory variables X, M and C (the explained portion) and that attributable to group differences in the effects of X,M and C on the mean outcome Y, as well as group differences in the mean

outcome at the reference levels of X, M, and C (i.e. due to differences in the generating functions, the unexplained portion). Fortin 2011[14] considers several interpretations of the unexplained portion of the aggregate decomposition under conditional exchangeability $Y(r) \coprod R|X, M, C$. Given this, the effect of assigning to blacks the generating function of whites has the causal interpretation as an effect of assigning black vs. white racial classification among blacks (average effect of treatment on the treated); the effect of assigning to whites the generating function of blacks has the causal interpretation as an effect of assigning black vs. white racial classification among whites (average effect of treatment on the untreated); what one would consider the effect of everyone having the generating function of blacks vs. that of whites has the causal interpretation as an effect of assigning black vs. white racial classification among both groups (average effect of treatment). These quantities will differ when the effects of C, X or M on Y vary by race. Nonetheless, as pointed out by Fortin 2011,[14] interpreting any unexplained portion as an effect of race with counterfactuals may be problematic on several grounds: (i) race is non-manipulable; (ii) there may be differential selection into the sample by race based on unobsvervable characteristics (unmeasured selection-bias[46]); and (iii) not all structural determinants of the outcome that are associated with race are observed (unmeasured confounding). Note that (iii) will arise if there are unmeasured common causes of race and the outcome (e.g. of R and Y), and also when there are unmeasured common causes of mediators and the outcome (e.g., of M and Y). This gets to the chief concern that the conditional exchangeability assumption rests on a post-treatment variable M affected by race. Identification would be problematic if there were unmeasured confounding of M. Such unmeasured confounding would induce a selection-bias between race and the outcome through unmeasured determinants of the outcome associated with M.[46] Even with no unmeasured confounding of M, what is identified is not a total effect of race but rather a direct effect of race with respect to M. Perhaps for this reason the unexplained portion is viewed as a pure measure of discrimination.

Suppose now that in our causal structure race is unconfounded and that there are two mediators L and M as in our extended example, for whom each of their individual effects on Y do not interact with one another to cause Y and that L does not affect M. Suppose also that the effect of race is unconfounded. Furthermore, suppose that the effects of L and M are unconfounded given race alone (and there is no interest in the explanatory value of baseline covariates X and C). Huber 2015[15] showed that, under this model, the explained/unexplained components from a standard, parametric Oaxaca-Blinder decomposition equal the natural indirect/direct effects[47] from a parametric product-method estimator for multiple mediators. As pointed out by Huber 2015,[15] even with allowing for confounding by X, C through conditional exchangeability assumptions,[48] and relaxing the no-interaction assumptions through a weighting estimator, this equivalence between Oaxaca-Blinder decompositions and natural direct/indirect effects disallows the existence of any other mediators that confound those in view. If such a confounder did exist, it would not be possible to identify the natural direct/indirect effects even if it were measured. In many settings it may be difficult to interpret the results of aggregate Oaxaca-Blinder decompositions as natural indirect /direct effects. Moreover, this framing maintains causal inference with respect to race/group status.

Let us return to our original graph where racial classification is a function of history with X and C present (and possibly L). In our contribution, we introduced conditional forms of the aggregate Oaxaca-Blinder decomposition that consider a set of variables, and parse them into explanatory variables of interest and conditioning variables used to control for confounding of the explanatory variables, wherein causal interpretation is made with respect to intervene to set the distribution of the explanatory variables and not race. We showed that, under the conditional exchangeability

assumption $E[Y(m)|r,x,c] = E[Y(m)|m,r,x,c]$, the explained portion of an Oaxaca-Blinder decomposition within levels of X and C, with M as the explanatory variable, is equivalent to the disparity reduction under a randomized intervention to set M according to its distribution among whites $R = 0$ given X and C, the result in VanderWeele & Robinson 2014 (Proposition 2). The unexplained portion also has a causal interpretation as the residual disparity under that intervention. We also showed that, under the conditional exchangeability assumptions $E[Y(x,m)|r,x,c] = E[Y(x,m)|m,r,x,c]$ and $E[Y(x,m)|x,r,c] = E[Y(x,m)|r,c]$ an Oaxaca-Blinder decomposition within levels of C, with X and M as explanatory variables, is equivalent to the disparity reduction under a randomized intervention to set X and M according to their distribution among whites $R = 0$ given C (Proposition 3). Again, the unexplained portion has a causal interpretation as the residual disparity under that intervention. We also generalized these results to the case where there is a time-dependent confounder L tof the $M - Y$ relationship that is affected by R and/or X but L itself is not of explanatory interest (Propositions 5 and 6).

Some further remarks about the aggregate decomposition are in order. First, the causal interpretation we develop here for the aggregate decomposition with respect to the covariates will in most cases only be possible if one can establish conditional exchangeability for each explanatory variable given those that precede it. This may be difficult to accomplish when there are many explanatory variables; it may be best to pursue specific hypotheses for a subset of manipulable explanatory variables where conditional exchangeability can be satisfied. Our second remark is that in some cases, variables C will be needed to achieve conditional exchangeability but an investigator may not be interested in a disparity that conditions upon some or all variables in C (similar to our proposition 4 that does not condition on X). In this case, our results suggest that one needs to make use of a detailed decomposition that we consider at the end of the next section.

Detailed Decomposition

The goal of the detailed decomposition in the Oaxaca-Blinder decomposition literature is to apportion each of the explained and unexplained portions to each explanatory variable of interest, in our context, X and M. Rothe 2015[6] considers a detailed decomposition of the explained portion for, in our context, racial differences in the distribution of the outcome Y using copula functions. The approach reduces to the standard detailed decomposition of the explained portion for racial differences in the mean outcome under linear models for the outcome. Under an assumption that is essentially equivalent to conditional exchangeability for each covariate given the others (e.g., $Y(x,m) \coprod x\,|r,m,c$ and $Y(x,m) \coprod m\,|r,x,c$), each component of the explained portion is interpreted counterfactually: what difference in the outcome would one observe under an intervention to set the distribution of one or more covariates among blacks such that they follow the distribution among whites, while holding the distribution of other covariates fixed? The causal interpretation here is still tenuous for most variables. For example, consider an intervention on X in a decomposition that involved X and M. The exchangeability condition would be $Y(x,m) \coprod x\,|r,m,c$ which requires identification using a post-treatment variable M. Such post-treatment variables may affect the outcome or be confounded with the outcome by unmeasured variables, where conditioning on them results in selection-bias.[46] Moreover, even if confounders for the post-treatment variable M are assumed to be measured as in our example, such that no selection-bias would ensue, including M in the model renders the policy interpretation for X difficult because X affects M. The policy interpretation would pertain to an intervention that changes the distribution

of X while disallowing any resulting changes in the distribution of M. When X and M are causally related, it is unclear what intervention on X would not also affect M.

We showed that, under the conditional exchangeability assumption $E[Y(m)|r,x,c] = E[Y(m)|m,r,x,c]$, the explained portion of a detailed Oaxaca-Blinder decomposition within levels of C, corresponding to explanation by differences in the distribution of M, is equivalent to the disparity reduction under a randomized intervention to set M according to its distribution among whites $R = 0$ given C (Proposition 4). We also generalized this result to the case where there is a time-dependent confounder L of the $M - Y$ relationship that is affected by R and/or X (Proposition 7). The portion corresponding to explanation by differences in the distribution of X only has an analogous causal interpretation when the effect of X on Y is not mediated by M, in which case it is equivalent to an intervention to set the distribution of X according to its distribution among whites $R = 0$ given C (proposition 1), provided conditional exchangeability $E[Y(x,m)|r,c] = E[Y(x,m)|x,r,c]$ and $E[Y(x,m)|m,x,r,c] = E[Y(x,m)|x,r,c]$. Of course, if one were interested in proposition 1, one could just carry out an aggregate decomposition with just X as an explanatory variable under the exchangeability condition $E[Y(x)|r,c] = E[Y(x)|x,r,c]$ without any assumptions about the causal relationship between X and M.

Our results show that some results of a detailed decomposition have clear causal interpretations (and policy implications) while others do not. When conditional exchangeability only rests on pre-treatment variables as in our approach, the component for the ultimate variable (M in our example) has a clear causal (and policy-relevant) interpretation. The component of the explained portion for the ultimate variable can be interpreted as the disparity reduced an intervention to equalize the distribution of the variable, marginally with respect to the preceding variables, but not to set race R. The remainder of the observed disparity can be interpreted as the residual disparity after such an intervention. Thus, our results for propositions 1 and 4 suggest if one is interested in the potential effects of intervening on each target for reducing disparities, one should carry out a separate Oaxaca-Blinder decomposition for each target of interest, only including as many other pre-target variables needed to satisfy conditional exchangeability for the target variable of interest.

Our results also have implications for when investigators wish to use covariates to control for confounding but do not want to condition the disparity on some or all of them. It follows from our non-parametric results on propositions 4 that if C were empty, and we had temporally ordered multivariate $\overline{\boldsymbol{X}}$ $(X_1,\ldots,X_k)$ and $\overline{\boldsymbol{M}}$ $(M_1,\ldots,M_k)$ such that X_{k-1} preceded X_k and likewise for M_{k-1} and M_k, one could in fact make inferences about how a marginal disparity would change under an intervention to assign the joint distribution of $\overline{\boldsymbol{M}}$ among blacks to its distribution among whites, marginally with respect to $\overline{\boldsymbol{X}}$, provided that for each M_k we had conditional exchangeability given $\boldsymbol{X}$ and $\boldsymbol{M}_{k-1}$. One could also make use of a weaker exchangeability conditions involving possibly multivariate sets of time-dependent confounders $\overline{\boldsymbol{L}}$ $(L_1,\ldots,L_k)$ under extensions of our non-parametric results for proposition 7. Overall, our results make clear that a careful consideration of temporality is paramount for carrying a causally meaningful detailed Oaxaca-Blinder decomposition.

Implications

Our contributions lay the ground work for causal decompositions in a design and analysis framework that seeks to satisfy identifiability conditions for a refined set of potentially manipulable targets, regardless of whether they are non-mediating variables, mediating variables, or affected by other mediating variables. This allows for investigators to pursue causally meaningful Oxaca-Blinder decompositions while defining a disparity, possibly conditional on baseline factors X or C, but considering an intervention that may condition on all, some, or no baseline variables, or even variables L affected by race. Thus, they allow for decompositions that can be tailored as needed to particular substantive settings.

Results for proportion of the disparity reduced

Let D equal the total disparity measured on the difference scale $E[Y|R=1,c]-E[Y|R=0,c]$
Let D* equal the residual disparity measured on the difference scale $\mu-E[Y|R=0,c]$
Let R equal the total disparity measured on the relative scale $E[Y|R=1,c]/E[Y|R=0,c]$
Let R* equal the residual disparity measured on the relative scale $\mu/E[Y|R=0,c]$

Using additive disparity measures

Proportion of disparity remaining = D^*/D
Proportion of disparity reduced = $(D-D^*)/D$

Using relative disparity measures

Proportion of disparity remaining = $(R^*-1)/(R-1)$
Proportion of disparity reduced = $(R-R^*)/(R-1)$

Appendix Table 1. Results under parametric regression models for a continuous outcome Y (in the absence of time-dependent confounding).		
	Successive linear models for Y $E[Y\|r,x,m,c]$ $= \theta_0 + \theta_1 r + \theta_2 x + \theta_3 m + \theta_4' c$ $E[Y\|r,x,c]$ $= \gamma_0 + \gamma_1 r + \gamma_2 x + \gamma_4' c$ $E[Y\|r,c]$ $= \phi_0 + \phi_1 r + \phi_4' c$	Linear models for Y, M, X $E[Y\|r,x,m,c]$ $= \theta_0 + \theta_1 r + \theta_2 x + \theta_3 m + \theta_4' c$ $E[M\|r,x,c]$ $= \beta_0 + \beta_1 r + \beta_2 x + \beta_3' c$ $E[X\|r,c]$ $= \alpha_0 + \alpha_1 r + \alpha_2' c$
Proposition 1		
Residual disparity[a]	γ_1	$\theta_1 + \beta_1\theta_3$
Disparity reduction[b]	$\phi_1 - \gamma_1$	$\alpha_1\theta_2 + \alpha_1\beta_2\theta_3$
Proposition 2		
Residual disparity[a]	θ_1	θ_1
Disparity reduction[b]	$\gamma_1 - \theta_1$	$\beta_1\theta_3$
Proposition 3		
Residual disparity[a]	θ_1	θ_1
Disparity reduction[b]	$\phi_1 - \theta_1$	$\alpha_1\theta_2 + \beta_1\theta_3 + \alpha_1\beta_2\theta_3$
Proposition 4		
Residual disparity[a]	$\theta_1 + (\theta_2/\gamma_2)(\phi_1 - \gamma_1)$	$\theta_1 + \alpha_1\theta_2$
Disparity reduction[b]	$(\gamma_1 - \theta_1) + (1-\theta_2/\gamma_2)(\phi_1 - \gamma_1)$	$\beta_1\theta_3 + \alpha_1\beta_2\theta_3$
[a] $\mu - E[Y\|R=0,c]$ [b] $E[Y\|R=1,c] - \mu$ where μ equals the mean counterfactual outcome for group R=1 under the proposed intervention		

Appendix Table 2. Results under parametric regression models for a rare binary outcome Y (in the absence of time-dependent confounding).		
	Successive logistic models for Y Logit $P[Y\|r,x,m,c]$ $= \theta_0 + \theta_1 r + \theta_2 x + \theta_3 m + \theta_4' c$ Logit $P[Y\|r,x,c]$ $= \gamma_0 + \gamma_1 r + \gamma_2 x + \gamma_4' c$ Logit $P[Y\|r,c]$ $= \phi_0 + \phi_1 r + \phi_4' c$	Models for Y, M, X Logit $P[Y\|r,x,m,c]$ $= \theta_0 + \theta_1 r + \theta_2 x + \theta_3 m + \theta_4' c$ $E[M\|r,x,c]$ $= \beta_0 + \beta_1 r + \beta_2 x + \beta_3' c$ $E[X\|r,c]$ $= \alpha_0 + \alpha_1 r + \alpha_2' c$
Proposition 1		
Residual disparity[a]	$\exp\{\gamma_1\}$	$\exp\{\theta_1 + \beta_1\theta_3\}$
Disparity reduction[b]	$\exp\{\phi_1 - \gamma_1\}$	$\exp\{\alpha_1\theta_2 + \alpha_1\beta_2\theta_3\}$
Proposition 2		
Residual disparity[a]	$\exp\{\theta_1\}$	$\exp\{\theta_1\}$
Disparity reduction[b]	$\exp\{\gamma_1 - \theta_1\}$	$\exp\{\beta_1\theta_3\}$
Proposition 3		
Residual disparity[a]	$\exp\{\theta_1\}$	$\exp\{\theta_1\}$
Disparity reduction[b]	$\exp\{\phi_1 - \theta_1\}$	$\exp\{\alpha_1\theta_2 + \beta_1\theta_3 + \alpha_1\beta_2\theta_3\}$
Proposition 4		
Residual disparity[a]	$\exp\{\theta_1 + (\theta_2/\gamma_2)(\phi_1 - \gamma_1)\}$	$\exp\{\theta_1 + \alpha_1\theta_2\}$
Disparity reduction[b]	$\exp\{(\gamma_1 - \theta_1) + (1-\theta_2/\gamma_2)(\phi_1 - \gamma_1)\}$	$\exp\{\beta_1\theta_3 + \alpha_1\beta_2\theta_3\}$
[a] $\mu - E[Y\|R=0,c]$ [b] $E[Y\|R=1,c] - \mu$ where μ equals the mean counterfactual outcome for group R=1 under the proposed intervention		

Results for successive linear models given measures of childhood characteristics, X_1, X_2, X_3

Consider the following models:

$E[Y|r,x_1,x_2,x_3,m,c] = \theta_0 + \theta_1 r + \theta_2 x_1 + \theta_3 x_2 + \theta_4 x_3 + \theta_5 m + \theta_6' c$

$E[Y|r,x_1,x_2,x_3,c] = \delta_0 + \delta_1 r + \delta_2 x_1 + \delta_3 x_2 + \delta_4 x_3 + \delta_6' c$

$E[Y|r,x_1,x_2,c] = \eta_0 + \eta_1 r + \eta_2 x_1 + \eta_3 x_2 + \eta_6' c$

$E[Y|r,x_1,c] = \gamma_0 + \gamma_1 r + \gamma_2 x_1 + \gamma_6' c$

$E[Y|r,c] = \phi_0 + \phi_1 r + \phi_6' c$

In Proposition 1 we have:

The residual disparity is: $\mu_{x1,x2,x3} - E[Y|R=0,c] = \delta_1$

The disparity reduction is: $E[Y|R=1,c] - \mu_{x1,x2,x3} = \phi_1 - \delta_1$

In Proposition 2 we have:

The residual disparity is: $\mu_{m|x1,x2,x3} - E[Y|R=0,x_1,x_2,x_3,c] = \theta_1$

The disparity reduction is: $E[Y|R=1,x_1,x_2,x_3,c] - \mu_{m|x1,x2,x3} = \delta_1 - \theta_1$

In Proposition 3 we have:

The residual disparity is: $\mu_{x1,x2,x3,m} - E[Y|R=0,c] = \theta_1$

The disparity reduction is: $E[Y|R=1,c] - \mu_{x1,x2,x3,m} = \phi_1 - \theta_1$

In Proposition 4 we have:

The residual disparity is:

$$\begin{aligned}
&\mu_m - E[Y|R=0,c] \\
&= \theta_1 \\
&+ \theta_4/\delta_4(\eta_1 - \delta_1) \\
&+ \{\theta_3/\eta_3 + \theta_4/\delta_4 (1 - \delta_3/\eta_3)\}(\gamma_1 - \eta_1) \\
&+ \{\theta_2/\gamma_2 + \theta_3/\eta_3(1-\eta_2/\gamma_2) + \theta_4/\delta_4\{(\eta_2 - \delta_2)/\gamma_2 + (1 - \delta_3/\eta_3)(1-\eta_2/\delta_2)\}\}(\phi_1 - \gamma_1)
\end{aligned}$$

The disparity reduction is:

$$\begin{aligned}
&E[Y|R=0,c] - \mu_m \\
&= (\delta_1 - \theta_1) \\
&+ (1-\theta_4/\delta_4)(\eta_1 - \delta_1) \\
&+ \{(\delta_3 - \theta_3)/\eta_3 + (1 - \theta_4/\delta_4)(1 - \delta_3/\eta_3)\}(\gamma_1 - \eta_1) \\
&+ \{(\delta_2 - \theta_2)/\gamma_2 + (\delta_3 - \theta_3)/\eta_3(1-\eta_2/\gamma_2) + (1-\theta_4/\delta_4)\{(\eta_2 - \delta_2)/\gamma_2 + (1 - \delta_3/\eta_3)(1-\eta_2/\delta_2)\}\}(\phi_1 - \gamma_1)
\end{aligned}$$

Appendix Table 3. Characteristics of males in the 1997 National Survey of American Youth Analytic Cohort, mean (standard deviation)		
	White (n=2,413)	Black (n=1,169)
Age	24.6 (1.5)	24.5 (1.5)
Adult outcomes		
Wage (dollars/hour)	21.3 (13.6)	17.3 (10.1)
Unemployed[a]	6.9 (2.5)	17.5 (3.8)
Incarceration, ever[a]	8.2 (2.7)	16.4 (3.7)
Educational Attainment		
Armed Forces Qualifying Test (AFQT; z-score)	0.35 (0.98)	-0.66 (0.78)
Total years education (years)	12.7 (1.8)	11.9 (1.4)
Measures of childhood SES		
Mother's highest grade level	13.5 (2.5)	12.5 (2.1)
Parental net worth in childhood (dollars)	$137,933 ($160,945)	$35,994 ($67,260)
Household Income in childhood (dollars)	$59,506 ($46,673)	$30,262 ($29,051)
Proportion missing (%)		
Missing AFQT[a]	17.7 (38.2)	24.8 (43.2)
Missing total years of education[a]	18.1 (38.5)	16.9 (37.5)
Missing mother's highest grade level[a]	8.1 (27.3)	14.6 (35.4)
Missing parental net worth in childhood[a]	25.4 (43.6)	28.0 (44.9)
Missing household income in childhood[a]	22.7 (41.9)	31.1 (46.3)
[a]Binary variable (1=yes, 0=no), scaled by 100. E.g., 6.9% of NLSY97 whites were unemployed in 2006.		

Appendix Table 4. Estimates of residual disparities and disparity reductions in adult outcomes under hypothetical intervention strategies on childhood SES measures and/or Armed Forces Qualifying Test scores in the 1997 NLSY Cohort[2]					
	Proposition 1	Proposition 2	Proposition 3	Proposition 4	Re-analysis of Fryer
	Intervene to equalize the distribution of childhood SES measures across race but not AFQT scores	Intervene to equalize the distribution of AFQT scores across race within levels of childhood SES	Intervene to equalize the distribution of AFQT scores and childhood SES measures across race	Intervene to equalize the distribution of AFQT scores across race but not childhood SES measures	Statistically equalize the distribution of AFQT scores across race without control for childhood SES
Log wages					
Initial disparity	-0.19 (0.02)	-0.14 (0.02)	-0.19 (0.02)	-0.19 (0.02)	-0.19 (0.02)
Residual disparity	-0.14 (0.02)	-0.10 (0.03)	-0.10 (0.03)	-0.13 (0.03)	-0.12 (0.03)
% reduction	25	34	51	32	38
Incarceration					
Initial disparity	2.12 (1.12)	1.65 (1.13)	2.12 (1.12)	2.12 (1.12)	2.12 (1.12)
Residual disparity	1.65 (1.13)	1.22 (1.13)	1.22 (1.13)	1.43 (1.13)	1.39 (1.13)
% reduction	54	34	18	36	32
Unemployment					
Initial disparity	2.86 (1.15)	2.39 (1.16)	2.86 (1.15)	2.86 (1.15)	2.86 (1.15)
Residual disparity	2.39 (1.16)	1.95 (1.17)	1.95 (1.17)	2.21 (1.17)	2.12 (1.16)
% reduction	26	31	49	35	40
[2]The analytic sample size was 3279 for wages, 3294 for unemployment, and 4599 for incarceration. All models included mutually exclusive dummy variables for Hispanic ethnicity and mixed race.					

Appendix Table 5. Estimates of residual disparities and disparity reductions in adult outcomes under hypothetical intervention strategies on childhood SES measures and/or total years of education in the 1997 NLSY Cohort[2]					
	Proposition 1	Proposition 2	Proposition 3	Proposition 4	Re-analysis of Fryer
	Intervene to equalize the distribution of childhood SES measures across race but not total years of education	Intervene to equalize the distribution of total years of education across race within levels of childhood SES	Intervene to equalize the distribution of total years of education and childhood SES measures across race	Intervene to equalize the distribution of total years of education across race but not childhood SES measures	Statistically equalize the distribution of total years of education across race without control for childhood SES
Log wages					
Initial disparity	-0.19 (0.02)	-0.14 (0.02)	-0.19 (0.02)	-0.19 (0.02)	-0.19 (0.02)
Residual disparity	-0.14 (0.02)	-0.13 (0.02)	-0.13 (0.02)	-0.16 (0.03)	-0.15 (0.02)
% reduction	25	11	33	19	21
Incarceration					
Initial disparity	2.22 (1.12)	1.66 (1.14)	2.22 (1.12)	2.22 (1.12)	2.22 (1.12)
Residual disparity	1.66 (1.14)	1.50 (1.14)	1.50 (1.14)	1.74 (1.42)	1.69 (1.13)
% reduction	46	24	59	41	43
Unemployment					
Initial disparity	2.86 (1.13)	2.39 (1.15)	2.86 (1.13)	2.86 (1.13)	2.86 (1.13)
Residual disparity	2.39 (1.15)	2.32 (1.15)	2.32 (1.15)	2.64 (1.58)	2.53 (1.14)
% reduction	26	6	30	15	18
[2]The analytic sample size was 3279 for wages, 3294 for unemployment, and 4599 for incarceration. All models included mutually exclusive dummy variables for Hispanic ethnicity and mixed race.					

Proofs

Non-parametric formulae in the absence of time-dependent confounding

Our assumptions are:
A1: The effect of X on the outcome Y is unconfounded given (R,C)
A2: The effect of M on the outcome Y is unconfounded given (R,C,X)
Along with consistency and positivity for X and M

Formally these are:

I. Conditional exchangeability:
A1: $E[Y(x)|R=r,c] = E[Y(x)|R=r,x,c]$
A1': $E[Y(x,m)|R=r,c] = E[Y(x,m)|R=r,x,c]$
A2: $E[Y(m)|R=r,x,c] = E[Y(m)|R=r,x,m,c]$
A2': $E[Y(x,m)|R=r,x,c] = E[Y(x,m)|R=r,x,m,c]$

II. Consistency (for individual i):
If $X_i=x_i$ then $Y_i(x)=Y_i$
If $M_i=m$ then $Y_i(m)=Y_i$
If $X_i=x$ and $M_i=m_i$ then $Y_i(x,m)=Y_i$

III. Positivity (common support among defined population of interest):
$f_{X|R,C}(x|r,c) > 0$ for all R, C *where* $P[R=1, C=c] > 0$ and $P[R=0, C=c] > 0$
$f_{M|R,X,C}(m|r,x,c) > 0$ for all R, X, C *where* $P[R=1, X=x, C=c] > 0$
and $P[R=0, X=x, C=c] > 0$

Recall Proposition 1 (VanderWeele and Robinson, 2014). The disparity that would remain if the childhood distribution of X for black persons (R=1) with covariates C=c were set equal to its distribution for white persons (R=0) with C=c would be:
$\mu_x - E[Y|R=0,c]$
and the amount the disparity is reduced would be:
$E[Y|R=1,c] - \mu_x$
where $\mu_x = \Sigma_x E[Y|R=1,x,c]P(x|R=0,c)$.

Proof of Proposition 1: Let $G_{x|c}$ denote a random draw of the distribution of X among those with R=0,C=c i.e. from P(x|R=0,c). If the distribution of X for black persons (R=1) with covariates C=c were set equal to its distribution for white persons (R=0) the average outcome would be:
$E[Y(x=G_{x|c})|R=1,c]$
$= \Sigma_x E[Y(x)|R=1,c, G_{x|c}=x]P(G_{x|c}=x \mid R=1,c)$
$= \Sigma_x E[Y(x)|R=1,c] P(x|R=0,c)$ by definition of $G_{x|c}=x$ as random given C=c
$= \Sigma_x E[Y(x)|R=1,x,c] P(x|R=0,c)$ by (A1)
$= \Sigma_x E[Y|R=1,x,c] P(x|R=0,c)$.
From this the result follows.

Recall Proposition 2 (VanderWeele and Robinson, 2014). The disparity that would remain if the distribution of M for black persons (R=1) with covariates C=c and X=x were set equal to its distribution for white persons (R=0) with C=c and X=x would be:
$\mu_{m|x}$ -E[Y|R=0,x,c]
and the amount the disparity is reduced would be:
E[Y|R=1,x,c]- $\mu_{m|x}$
where $\mu_{m|x} = \Sigma_m$ E[Y|R=1,x,m,c]P(m|R=0,x,c).

Proof of Proposition 2. Let $G_{m|x,c}$ denote a random draw of the distribution of M among those with R=0,C=c,X=x i.e. from P(m|R=0,x,c). If the distribution of M for black persons (R=1) with covariates C=c and X=x were set equal to its distribution for white persons (R=0) with covariates C=c and X=x would be:
E[Y(m)=$G_{m|x,c}$)|R=1,x,c]
$= \Sigma_m$ E[Y(m)|R=1,x,c, $G_{m|x,c}$=m] P($G_{m|x,c}$ = m | R=1,x,c)
$= \Sigma_m$ E[Y(m)|R=1,x,c] P(m|R=0,x,c) by definition of $G_{m|x,c}$=m as random given C=c
$= \Sigma_m$ E[Y(m)|R=1,x,m,c] P(m|R=0,x,c) by (A2)
$= \Sigma_m$ E[Y|R=1,x,m,c] P(m|R=0,x,c) by (A2)
From this the result follows.

Recall Proposition 3. The disparity that would remain if the distribution of (X,M) for black persons (R=1) with covariates C=c were set equal to its distribution for white persons (R=0) with C=c would be:
μ_{xm} -E[Y|R=0,c]
and the amount the disparity is reduced would be:
E[Y|R=1,c]- μ_{xm}
where $\mu_{xm} = \Sigma_{x,m}$ E[Y|R=1,x,m,c]P(m|R=0,x,c)P(x|R=0,c).

Proof of Proposition 3. Let $G_{xm|c}$ denote a random draw of the distribution of (X,M) among those with R=0,C=c i.e. from P(m,x|R=0,c). If the distribution of (X,M) for black persons (R=1) with covariates C=c were set equal to its distribution for white persons (R=0) with covariates C=c would the average outcome would be:
E[Y((x,m)=$G_{x,m|c}$)|R=1,c]
$= \Sigma_{x,m}$ E[Y(x,m)|R=1,c, $G_{x,m|c}$ =(x,m)] P($G_{x,m|c}$ =(x,m)| R=1,c)
$= \Sigma_{x,m}$ E[Y(x,m)|R=1,c] P(m,x|R=0,c) by definition of $G_{x,m|c}$=x,m as random given C=c
$= \Sigma_{x,m}$ E[Y(x,m)|R=1,x,c] P(m|R=0,x,c)P(x|R=0,c) by (A1')
$= \Sigma_{x,m}$ E[Y(x,m)|R=1,x,m,c] P(m|R=0,x,c)P(x|R=0,c) by (A2')
$= \Sigma_{x,m}$ E[Y|R=1,x,m,c] P(m|R=0,x,c)P(x|R=0,c).
From this the result follows.

Recall Proposition 4. The disparity that would remain if the distribution of M for black persons (R=1) with covariates C=c were set equal to its distribution for white persons (R=0) with C=c would be:
μ_m -E[Y|R=0,c]
and the amount the disparity is reduced would be:
E[Y|R=1,c]- μ_m
where $\mu_m = \Sigma_{x,m}$ E[Y|R=1,x,m,c]P(m|R=0,c)P(x|R=1,c).

Proof of Proposition 4. Let $G_{m|c}$ denote a random draw of the distribution of M among those with R=0,C=c i.e. from P(m|R=0,c). If the distribution of M for black persons (R=1) with covariates C=c were set equal to its distribution for white persons (R=0) with covariates C=c the average outcome would be:
$E[Y(m)=G_{m|c})|R=1,c]$
$= \Sigma_m E[Y(m)|R=1,c, G_{m|c}=m] P(G_{m|c} = m \mid R=1,c)$
$= \Sigma_m E[Y(m)|R=1,c] P(m|R=0,c)$
$= \Sigma_{x,m} E[Y(m)|R=1,x,c] P(x|R=1,c) P(m|R=0,c)$
$= \Sigma_{x,m} E[Y(m)|R=1,x,m,c] P(x|R=1,c) P(m|R=0,c)$ by (A2)
$= \Sigma_{x,m} E[Y|R=1,x,m,c] P(x|R=1,c) P(m|R=0,c)$.
From this the result follows.

Non-parametric formulae in the presence of time-dependent confounding

Suppose now that there is a variable L, that may be affected by C,R,X and that affects both M and Y so that it is a confounder of the relationship between M and Y.

We will assume:
A1: The effect of X on the outcome Y is unconfounded given (R,C)
A3: The effect of M on the outcome Y is unconfounded given (R,C,X,L)
Along with positivity and consistency for X and M

Formally these are:
I. Conditional exchangeability:
A1: $E[Y(x)|R=r,c] = E[Y(x)|R=r,x,c]$
A3: $E[Y(m)|R=r,x,c,l] = E[Y(m)|R=r,x,m,c,l]$
A3': $E[Y(x,m)|R=r,x,c,l] = E[Y(x,m)|R=r,x,m,c,l]$

II. Consistency (for individual i):
If $X_i=x_i$ then $Y_i^x=Y_i$
If $M_i=m$ then $Y_i^m=Y_i$
If $X_i=x$ and $M_i=m_i$ then $Y_i^{x,m}=Y_i$

III. Positivity (common support among defined population of interest):
$f_{X|R,C}(x|r,c) > 0 \; for \; all \; R, C \; where \; P[R=1, C=c] > 0$ and $P[R=0, C=c] > 0$
$f_{M|R,L,X,C}(m|r,l,x,c) > 0$ for all R, L, X, C where
$P[R=1, L=l, X=x, C=c] > 0$ and $P[R=0, L=l, X=x, C=c] > 0$

Recall Proposition 5. Under (A3), the disparity that would remain if the distribution of M for black persons (R=1) with X=x and covariates C=c were set equal to its distribution for white persons (R=0) with X=x and C=c would be:
$\mu_{m|x}$ -E[Y|R=0,x,c]
and the amount the disparity is reduced would be:
E[Y|R=1,x,c]- $\mu_{m|x}$
where $\mu_{m|x} = \Sigma_{m,l|x} E[Y|R=1,x,m,c,l]P(l|R=1,x,c)P(m|R=0,x,c)$.

Proof of Proposition 5. Let $G_{m|x,c}$ denote a random draw of the distribution of M among those with R=0,X=x,C=c i.e. from P(m|R=0,x,c). If the distribution of M for black persons (R=1) with childhood SES X=x and covariates C=c were set equal to its distribution for white persons (R=0) with childhood SES X=x and covariates C=c would the average outcome would be:
$E[Y((m)=G_{m|x,c})|R=1,x,c]$
$= \Sigma_m E[Y(m)|R=1,x,c,G_{m|x,c} =(m)] P(G_{m|x,c} =(m)| R=1,x,c)$
$= \Sigma_m E[Y(m)|R=1,x,c] P(m|R=0,x,c)$ by definition of $G_{m|x,c}$=x as random given C=c
$= \Sigma_{m,l} E[Y(m)|R=1,x,c,l] P(l|R=1,x,c)P(m|R=0,x,c)$
$= \Sigma_{m,l} E[Y(m)|R=1,x,m,c,l] P(l|R=1,x,c) P(m|R=0,x,c)$ by (A3)
$= \Sigma_{m,l} E[Y|R=1,x,m,c,l] P(l|R=1,x,c)P(m|R=0,x,c)$ by consistency.
From this the result follows.

Recall Proposition 6. Under (A1') and (A3'), the disparity that would remain if the distribution of (X,M) for black persons (R=1) with covariates C=c were set equal to its distribution for white persons (R=0) with C=c would be:
μ_{xm} -E[Y|R=0,c]
and the amount the disparity is reduced would be:
$E[Y|R=1,c]- \mu_{xm}$
where $\mu_{xm} = \Sigma_{x,m,l} E[Y|R=1,x,m,c,l]P(l|R=1,x,c)P(m|R=0,x,c)P(x|R=0,c)$.

Proof of Proposition 6. Let $G_{xm|c}$ denote a random draw of the distribution of (X,M) among those with R=0,C=c i.e. from P(m,x|R=0,c). If the distribution of (X,M) for black persons (R=1) with covariates C=c were set equal to its distribution for white persons (R=0) with covariates C=c would the average outcome would be:
$E[Y((x,m)=G_{xm|c})|R=1,c]$
$= \Sigma_{x,m} E[Y(x,m)|R=1,c, G_{xm|c} =(x,m)] P(G_{xm|c} =(x,m)| R=1,c)$
$= \Sigma_{x,m} E[Y(x,m)|R=1,c] P(m,x|R=0,c)$ by definition of $G_{m|x,c}$=x,m as random given C=c
$= \Sigma_{x,m} E[Y(x,m)|R=1,x,c] P(m|R=0,x,c)P(x|R=0,c)$ by (A1')
$= \Sigma_{x,m,l} E[Y(x,m)|R=1,x,c,l] P(l|R=1,x,c)P(m|R=0,x,c)P(x|R=0,c)$
$= \Sigma_{x,m,l} E[Y(x,m)|R=1,x,m,c,l] P(l|R=1,x,c) P(m|R=0,x,c)P(x|R=0,c)$ by (A3')
$= \Sigma_{x,m,l} E[Y|R=1,x,m,c,l] P(l|R=1,x,c)P(m|R=0,x,c)P(x|R=0,c)$ by consistency.
From this the result follows.

Recall Proposition 7. Under (A3), the disparity that would remain if the distribution of M for black persons (R=1) with covariates C=c were set equal to its distribution for white persons (R=0) with C=c would be:
μ_m -E[Y|R=0,c]
and the amount the disparity is reduced would be:
$E[Y|R=1,c]- \mu_m$
where $\mu_m = \Sigma_{x,m,l} E[Y|R=1,x,m,c,l]P(l|R=1,x,c)P(m|R=0,c)P(x|R=1,c)$.

Proof of Proposition 7. Let $G_{m|c}$ denote a random draw of the distribution of M among those with R=0,C=c i.e. from P(m|R=0,c). If the distribution of M for black persons (R=1) with covariates C=c were set equal to its distribution for white persons (R=0) with covariates C=c the average outcome would be:
$E[Y(m)=G_{m|c})|R=1,c]$
$= \Sigma_m E[Y(m)|R=1,c, G_{m|c}=m] P(G_{m|c} = m | R=1,c)$
$= \Sigma_m E[Y(m)|R=1,c] P(m|R=0,c)$ by definition of $G_{m|c}$=m as random given C=c
$= \Sigma_{x,m} E[Y(m)|R=1,x,c] P(x|R=1,c) P(m|R=0,c)$

$= \Sigma_{x,m,l}\, E[Y(m)|R=1,x,c,l]\, P(l|R=1,x,c)P(x|R=1,c)\, P(m|R=0,c)$
$= \Sigma_{x,m,l}\, E[Y(m)|R=1,x,m,c,l]\, P(l|R=1,x,c)P(x|R=1,c)\, P(m|R=0,c)$ by (A3)
$= \Sigma_{x,m,l}\, E[Y|R=1,x,m,c,l]\, P(l|R=1,x,c)\, P(x|R=1,c)\, P(m|R=0,c)$ by consistency.
From this the result follows.

Successive linear models for Y

(under a single measure of X)

Consider the following models:
$E[Y|r,x,m,c] = \theta_0 + \theta_1 r + \theta_2 x + \theta_3 m + \theta_4' c$
$E[Y|r,x,c] = \gamma_0 + \gamma_1 r + \gamma_2 x + \gamma_4' c$
$E[Y|r,c] = \phi_0 + \phi_1 r + \phi_4' c$

The results under the linear models for Propositions 1 and 2 were shown in VanderWeele and Robinson (2014).

The results under linear models for Proposition 3, to set the distribution of childhood SES and test scores (X,M) among black persons to their distribution among white persons, follow since:
$\mu_{xm} = \Sigma_{x,m}\, E[Y|R=1,x,m,c]\, P(m|R=0,x,c)P(x|R=0,c).$
$= \Sigma_{x,m}\, (\theta_0 + \theta_1 + \theta_2 x + \theta_3 m + \theta_4' c)\, P(m|R=0,x,c)P(x|R=0,c)$
$= \theta_0 + \theta_1 + \theta_2\, E[X|R=0,c] + \theta_3 E[M|R=0,c] + \theta_4' c$
Similarly,
$E[Y|R=0,c] = E[Y|R=0,x,m,c]\, P(m|R=0,x,c)P(x|R=0,c).$
$= \theta_0 + \theta_2\, E[X|R=0,c] + \theta_3 E[M|R=0,c] + \theta_4' c$
Thus, $\mu_{xm} - E[Y|R=0,c] = \theta_1$
Moreover,
$E[Y|R=1,c] - \mu_{xm} = \{E[Y|R=1,c] - E[Y|R=0,c]\} - \{\mu_{xm} - E[Y|R=0,c]\} = \phi_1 - \theta_1$

The results under linear models for Proposition 4, to set the distribution of test scores M among black persons to its distribution among white persons, follow since:

$\mu_m = \Sigma_{x,m}\, E[Y|R=1,x,m,c]P(m|R=0,c)P(x|R=1,c)$
$= \Sigma_{x,m}\, (\theta_0 + \theta_1 + \theta_2 x + \theta_3 m + \theta_4' c)P(m|R=0,c)P(x|R=1,c)$
$= \theta_0 + \theta_1 + \theta_2 E[X|R=1,c] + \theta_3 E[M|R=0,c] + \theta_4' c$
Similarly,
$E[Y|R=0,c] = \Sigma_{x,m}\, E[Y|R=0,x,m,c]\, P(m|R=0,x,c)P(x|R=0,c)$
$= \Sigma_{x,m}\, (\theta_0 + \theta_2 x + \theta_3 m + \theta_4' c)P(m|R=0,x,c)P(x|R=0,c)$
$= \theta_0 + \theta_2\, E[X|R=0,c] + \theta_3 E[M|R=0,c] + \theta_4' c$
Thus, $\mu_m - E[Y|R=0,c] = \theta_1 + \theta_2\, \{E[X|R=1,c] - E[X|R=0,c]\}$

Note that:
$E[Y|R=1,c] - E[Y|R=0,c] = \phi_1$
Also:
$E[Y|R=1,c] - E[Y|R=0,c]$
$= \Sigma_x\, E[Y|R=1,x,c]P(x|R=1,c) - \Sigma_x\, E[Y|R=0,x,c]P(x|R=0,c)$
$= \Sigma_x\, (\gamma_0 + \gamma_1 + \gamma_2 x + \gamma_4' c)P(x|R=1,c) - \Sigma_x\, (\gamma_0 + \gamma_2 x + \gamma_4' c)P(x|R=0,c)$

$= \gamma_1 + \gamma_2 \{E[X|R=1,c] - E[X|R=0,c]\}$
Thus: $\phi_1 = \gamma_1 + \gamma_2 \{E[X|R=1,c] - E[X|R=0,c]\}$
And so: $\{E[X|R=1,c] - E[X|R=0,c]\} = (\phi_1 - \gamma_1)/\gamma_2$

Therefore the remaining disparity is:
$\mu_m - E[Y|R=0,c]$
$= \theta_1 + \theta_2 \{E[X|R=1,c] - E[X|R=0,c]\}$
$= \theta_1 + \theta_2 (\phi_1 - \gamma_1)/\gamma_2$
And the disparity reduction is:
$E[X|R=1,c] - \mu_m = \{E[X|R=1,c] - E[X|R=0,c]\} - \{\mu_m - E[X|R=0,c\}$
$= \gamma_1 + \gamma_2 \{E[X|R=1,c] - E[X|R=0,c]\} - \theta_1 - \theta_2 \{E[X|R=1,c] - E[X|R=0,c]\}$
$= (\gamma_1 - \theta_1) + (\theta_2/\gamma_2)(\phi_1 - \gamma_1)$

Successive linear models for Y

(under multiple measures X i.e. X_1, X_2, X_3)

Suppose there were three potentially manipulable measures of early life characteristics X_1, X_2, X_3 as used in the motivating example. It can be shown that the proofs and non-parametric results above regarding propositions 1-4 apply replacing X with X_1, X_2, X_3 and x with x_1, x_2, x_3. Below we provide results under successive linear models for outcome Y, however it can be shown that the results also apply on the logit scale under successive logistic models for a rare binary outcome Y.

Consider the following linear models:
$E[Y|r,x_1,x_2,x_3,m,c] = \theta_0 + \theta_1 r + \theta_2 x_1 + \theta_3 x_2 + \theta_4 x_3 + \theta_5 m + \theta_6' c$
$E[Y|r,x_1,x_2,x_3,c] = \delta_0 + \delta_1 r + \delta_2 x_1 + \delta_3 x_2 + \delta_4 x_3 + \delta_6' c$
$E[Y|r,x_1,x_2,c] = \eta_0 + \eta_1 r + \eta_2 x_1 + \eta_3 x_2 + \eta_6' c$
$E[Y|r,x_1,c] = \gamma_0 + \gamma_1 r + \gamma_2 x_1 + \gamma_6' c$
$E[Y|r,c] = \phi_0 + \phi_1 r + \phi_6' c$

The results under linear models for proposition 1, to set the distribution of childhood SES X among black persons to its distribution among white persons, follow since:

$\mu_{x1,x2,x3} = \Sigma_{x1,x2,x3} E[Y|R=1,x_1,x_2,x_3,c]P(x_1,x_2,x_3|R=0,c)$
$= \Sigma_{x1,x2,x3} E[Y|R=1,x_1,x_2,x_3,c]P(x_3|R=0,x_1,x_2,c)P(x_2|R=0,x_1,c)P(x_1|R=0,c)$
$= \Sigma_{x1,x2,x3} (\delta_0 + \delta_1 + \delta_2 x_1 + \delta_3 x_2 + \delta_4 x_3 + \delta_6' c)P(x_3|R=0,x_1,x_2,c)P(x_2|R=0,x_1,c)P(x_1|R=0,c)$
$= \delta_0 + \delta_1 + \delta_2 E[X_1|R=0,c] + \delta_3 E[X_2|R=0,c] + \delta_4 E[X_3|R=0,c] + \delta_6' c$
Similarly,
$E[Y|R=0,c] = \Sigma_{x1,x2,x3} E[Y|R=0,x_1,x_2,x_3,c]P(x_1,x_2,x_3|R=0,c)$
$= \Sigma_{x1,x2,x3} E[Y|R=0,x_1,x_2,x_3,c]P(x_3|R=0,x_1,x_2,c)P(x_2|R=0,x_1,c)P(x_1|R=0,c)$
$= \Sigma_{x1,x2,x3} (\delta_0 + \delta_2 x_1 + \delta_3 x_2 + \delta_4 x_3 + \delta_6' c)P(x_3|R=0,x_1,x_2,c)P(x_2|R=0,x_1,c)P(x_1|R=0,c)$
$= \delta_0 + \delta_1 + \delta_2 E[X_1|R=0,c] + \delta_3 E[X_2|R=0,c] + \delta_4 E[X_3|R=0,c] + \delta_6' c$
Thus,
$\mu_{x1,x2,x3} - E[Y|R=0,c] = \delta_1$
Moreover,
$E[Y|R=1,c] - \mu_{x1,x2,x3} = \{E[Y|R=1,c] - E[Y|R=0,c]\} - \{\mu_{x1,x2,x3} - E[Y|R=0,c]\} = \phi_1 - \delta_1$

The results under linear models for proposition 2, to set the distribution of test scores M among black persons with childhood SES X=x to its distribution among white persons with childhood SES X=x, follow since:

$\mu_{m|x1,x2,x3} = \Sigma_m E[Y|R=1,m,x_1,x_2,x_3,c]P(m|R=0,x_1,x_2,x_3,c)$
$= \Sigma_m (\theta_0 + \theta_1 + \theta_2x_1 + \theta_3x_2 + \theta_4x_3 + \theta_5m + \theta_6'c)P(m|R=0,x_1,x_2,x_3,c)$
$= \theta_0 + \theta_1 + \theta_2x_1 + \theta_3x_2 + \theta_4x_3 + \theta_2E[M|R=0,x_1,x_2,x_3,c] + \theta_6'c$
Similarly,
$E[Y|R=0,x_1,x_2,x_3,c] = \Sigma_m E[Y|R=0,m,x_1,x_2,x_3,c]P(m|R=0,x_1,x_2,x_3,c)$
$= \Sigma_m (\theta_0 + \theta_2x_1 + \theta_3x_2 + \theta_4x_3 + \theta_5m + \theta_6'c)P(m|R=0,x_1,x_2,x_3,c)$
$= \theta_0 + \theta_2x_1 + \theta_3x_2 + \theta_4x_3 + \theta_2E[M|R=0,x_1,x_2,x_3,c] + \theta_6'c$
Thus,
$\mu_{m|x1,x2,x3} - E[Y|R=0,x_1,x_2,x_3,c] = \theta_1$
Moreover,
$E[Y|R=1,x_1,x_2,x_3,c]-E[Y|R=0,x_1,x_2,x_3,c]=\delta_1$
And so,
$E[Y|R=1,x_1,x_2,x_3,c]- \mu_{m|x1,x2,x3}$
$= \{E[Y|R=1,x_1,x_2,x_3,c]- E[Y|R=0,x_1,x_2,x_3,c]\} - \{\mu_{x1,x2,x3} - E[Y|R=0,c]\} = \delta_1-\theta_1$

The results under linear models for proposition 3, to set the distribution of childhood SES and test scores (X,M) among black persons to its distribution among white persons, follow since:

$\mu_{x1,x2,x3,m} = \Sigma_{x1,x2,x3,m} E[Y|R=1,x_1,x_2,x_3,m,c]P(m|R=0,x_1,x_2,x_3,c)P(x_1,x_2,x_3|R=0,c)$
$= \Sigma_{x1,x2,x3,m} E[Y|R=1,x_1,x_2,x_3,m,c]P(m|R=0,x_1,x_2,x_3,c)P(x_1,x_2,x_3|R=0,c)$
$= \Sigma_{x1,x2,x3,m} E[Y|R=1,m,x_1,x_2,x_3,c]\ P(m|R=0,x_1,x_2,x_3,c)\ P(x_3|R=0,x_1,x_2,c)\ P(x_2|R=0,x_1,c)\ P(x_1|R=0,c)$
$= \Sigma_{x1,x2,x3,m} (\theta_0 + \theta_2x_1 + \theta_3x_2 + \theta_4x_3 + \theta_5m + \theta_6'c)\ P(m|R=0,x_1,x_2,x_3,c)\ P(x_3|R=0,x_1,x_2,c)$
$P(x_2|R=0,x_1,c)\ P(x_1|R=0,c)$
$= \theta_0 + \theta_1 + \theta_2\ E[X_1|R=0,c] + \theta_3\ E[X_2|R=0,c] + \theta_4\ E[X_3|R=0,c] + \theta_5E[M|R=0,c] + \theta_6'c$
Similarly,
$E[Y|R=0,c] = \Sigma_{x1,x2,x3,m} E[Y|R=0,x_1,x_2,x_3,m,c]P(m|R=0,x_1,x_2,x_3,c)P(x_1,x_2,x_3|R=0,c)$
$= \theta_0 + \theta_2E[X_1|R=0,c] + \theta_3E[X_2|R=0,c] + \theta_4E[X_3|R=0,c] + \theta_5E[M|R=0,c] + \theta_6'c$
Thus, $\mu_{x1,x2,x3,m} - E[Y|R=0,c] = \theta_1$
Moreover,
$E[Y|R=1,c]- \mu_{x1,x2,x3,m} = \{E[Y|R=1,c]- E[Y|R=0,c]\} - \{\mu_{x1,x2,x3,m} - E[Y|R=0,c]\} = \phi_1 - \theta_1$

The results follow under linear models for proposition 4, to set the distribution of test scores M among black persons to its distribution among white persons, follow since:

$\mu_m = \Sigma_{x1,x2,x3,m} E[Y|R=1,x_1,x_2,x_3,m,c]P(m|R=0,c)P(x_1,x_2,x_3|R=1,c)$
$= \Sigma_{x1,x2,x3,m} E[Y|R=1,x_1,x_2,x_3,m,c]P(m|R=0,c)P(x_1,x_2,x_3|R=1,c)$
$= \Sigma_{x1,x2,x3,m} E[Y|R=1,m,x_1,x_2,x_3,c]\ P(m|R=0,c)\ P(x_3|R=1,x_1,x_2,c)\ P(x_2|R=1,x_1,c)\ P(x_1|R=1,c)$
$= \Sigma_{x1,x2,x3,m} (\theta_0 + \theta_2x_1 + \theta_3x_2 + \theta_4x_3 + \theta_5m + \theta_6'c)\ P(m|R=0,c)\ P(x_3|R=1,x_1,x_2,c)\ P(x_2|R=1,x_1,c)$
$P(x_1|R=0,c)$
$= \theta_0 + \theta_1 + \theta_2\ E[X_1|R=1,c] + \theta_3\ E[X_2|R=1,c] + \theta_4\ E[X_3|R=1,c] + \theta_5E[M|R=0,c] + \theta_6'c$
Similarly,
$E[Y|R=0,c] = \Sigma_{x1,x2,x3,m} E[Y|R=0,x_1,x_2,x_3,m,c]P(m|R=0,x_1,x_2,x_3,c)P(x_1,x_2,x_3|R=0,c)$
$= \theta_0 + \theta_2E[X_1|R=0,c] + \theta_3E[X_2|R=0,c] + \theta_4E[X_3|R=0,c] + \theta_5E[M|R=0,c] + \theta_6'c$
And so
$\mu_m - E[Y|R=0,c]$

$= \theta_1 + \theta_2\{E[X_1|R=1,c]-E[X_1|R=0,c]\} + \theta_3\{E[X_2|R=1,c]-E[X_2|R=0,c]\} + \theta_4\{E[X_3|R=1,c]-E[X_3|R=0,c]\}$

Note that:
$E[Y|R=1,c]-E[Y|R=0,c]= \phi_1$
Also:
$E[Y|R=1,c]-E[Y|R=0,c]$
$= \Sigma_{x1,x2,x3} E[Y|R=1,x_1,x_2,x_3,c]P(x_1,x_2,x_3|R=1,c)- \Sigma_{x1,x2,x3} E[Y|R=0,x_1,x_2,x_3,c]P(x_1,x_2,x_3|R=0,c)$
$= \Sigma_{x1,x2,x3} (\delta_0 + \delta_1 + \delta_2 x_1 + \delta_3 x_2 + \delta_4 x_3 + \delta_6'c)\ P(x_3|R=1,x_1,x_2,c)P(x_2|R=1,x_1,c)\ P(x_1|R=1,c)$
$- \Sigma_{x1,x2,x3} (\delta_0 + \delta_2 x_1 + \delta_3 x_2 + \delta_4 x_3 + \delta_6'c)P(x_3|R=0,x_1,x_2,c)\ P(x_2|R=0,x_1,c)\ P(x_1|R=0,c)$
$= \delta_1 + \delta_2\{E[X_1|R=1,c]-E[X_1|R=0,c]\} + \delta_3\{E[X_2|R=1,c]-E[X_2|R=0,c]\} + \delta_4\{E[X_3|R=1,c]-E[X_3|R=0,c]\}$
Also:
$E[Y|R=1,c]-E[Y|R=0,c]$
$= \Sigma_{x1,x2} E[Y|R=1,x_1,x_2,c]P(x_1,x_2|R=1,c)- \Sigma_{x1,x2} E[Y|R=0,x_1,x_2,c]P(x_1,x_2|R=0,c)$
$= \Sigma_{x1,x2} (\eta_0 + \eta_1 + \eta_2 x_1 + \eta_3 x_2 + \eta_6'c)\ P(x_2|R=1,x_1,c)\ P(x_1|R=1,c)$
$- \Sigma_{x1,x2} (\eta_0 + \eta_2 x_1 + \eta_3 x_2 + \eta_6'c)\ P(x_2|R=0,x_1,c)\ P(x_1|R=0,c)$
$= \eta_1 + \eta_2\{E[X_1|R=1,c]-E[X_1|R=0,c]\} + \eta_3\{E[X_2|R=1,c]-E[X_2|R=0,c]\}$
Also:
$E[Y|R=1,c]-E[Y|R=0,c]$
$= \Sigma_{x1} E[Y|R=1,x_1,c]P(x_1|R=1,c)- \Sigma_{x1} E[Y|R=0,x_1,c]P(x_1|R=0,c)$
$= \Sigma_{x1}(\gamma_0 + \gamma_1 + \gamma_2 x_1 + \gamma_6'c)\ P(x_1|R=1,c) - \Sigma_{x1} (\eta_0 + \eta_2 x_1 + \eta_6'c)\ P(x_1|R=0,c)$
$= \gamma_1 + \gamma_2\{E[X_1|R=1,c]-E[X_1|R=0,c]\}$
Thus:
$E[X_1|R=1,c]-E[X_1|R=0,c]=(\phi_1 - \gamma_1)/\gamma_2$
$E[X_2|R=1,c]-E[X_2|R=0,c]=\{(\gamma_1 - \eta_1)+(1-\eta_2/\gamma_2)\ (\phi_1 - \gamma_1)\}/\eta_3$
$E[X_3|R=1,c]-E[X_3|R=0,c]=\{(\eta_1 - \delta_1)+(\eta_2 - \delta_2)/\gamma_2\ (\phi_1 - \gamma_1)+(1- \delta_3/\eta_3)\{(\gamma_1 - \eta_1)+(1-\eta_2/\delta_2)(\phi_1 - \gamma_1)\}\}/\delta_4$

Thus, the residual disparity
$\mu_m - E[Y|R=0,c]$
$= \theta_1 + \theta_2\{E[X_1|R=1,c]-E[X_1|R=0,c]\} + \theta_3\{E[X_2|R=1,c]-E[X_2|R=0,c]\} + \theta_4\{E[X_3|R=1,c]-E[X_3|R=0,c]\}$
$= \theta_1$
$+ \theta_2/\gamma_2\ (\phi_1 - \gamma_1)$
$+ \theta_3/\eta_3\{(\gamma_1 - \eta_1)+ (1-\eta_2/\gamma_2)\ (\phi_1 - \gamma_1)\}$
$+ \theta_4/\delta_4\{(\eta_1 - \delta_1)+(\eta_2 - \delta_2)/\gamma_2\ (\phi_1 - \gamma_1)+(1- \delta_3/\eta_3)\{(\gamma_1 - \eta_1)+(1-\eta_2/\delta_2)(\phi_1 - \gamma_1)\}\}$
$= \theta_1$
$+ \theta_4/\delta_4(\eta_1 - \delta_1)$
$+ \{\theta_3/\eta_3 + \theta_4/\delta_4\ (1- \delta_3/\eta_3)\}(\gamma_1 - \eta_1)$
$+ \{\theta_2/\gamma_2 + \theta_3/\eta_3(1-\eta_2/\gamma_2) + \theta_4/\delta_4\{(\eta_2 - \delta_2)/\gamma_2 + (1- \delta_3/\eta_3)(\ 1-\eta_2/\delta_2)\}\}(\phi_1 - \gamma_1)$

And the disparity reduced
$E[Y|R=0,c]- \mu_m$
$=\{ E[Y|R=1,c]-E[Y|R=0,c]\}-\{\mu_m - E[Y|R=0,c]\}$
$= \delta_1 + \delta_2\{E[X_1|R=1,c]-E[X_1|R=0,c]\} + \delta_3\{E[X_2|R=1,c]-E[X_2|R=0,c]\} + \delta_4\{E[X_3|R=1,c]-E[X_3|R=0,c]\}$
$-\theta_1 - \theta_2\{E[X_1|R=1,c]-E[X_1|R=0,c]\} - \theta_3\{E[X_2|R=1,c]-E[X_2|R=0,c]\} - \theta_4\{E[X_3|R=1,c]-E[X_3|R=0,c]\}$
$= (\delta_1 - \theta_1) + (\delta_2-\theta_2)\{E[X_1|R=1,c]-E[X_1|R=0,c]\} + (\delta_3-\theta_3)\{E[X_2|R=1,c]-E[X_2|R=0,c]\}$
$+ (\delta_4-\theta_4)\{E[X_3|R=1,c]-E[X_3|R=0,c]\}$
$= (\delta_1 - \theta_1)$
$+ (\delta_2 - \theta_2)/\gamma_2\ (\phi_1 - \gamma_1)$

$+ (\delta_3 - \theta_3)/\eta_3\{(\gamma_1 - \eta_1) + (1-\eta_2/\gamma_2)\ (\phi_1 - \gamma_1)\}$
$+ (1-\theta_4/\delta_4)\{(\eta_1 - \delta_1) + (\eta_2 - \delta_2)/\gamma_2\ (\phi_1 - \gamma_1) + (1- \delta_3/\eta_3)\{(\gamma_1 - \eta_1) + (1-\eta_2/\delta_2)(\phi_1 - \gamma_1)\}\}$
$= (\delta_1 - \theta_1)$
$+ (1-\theta_4/\delta_4)(\eta_1 - \delta_1)$
$+ \{(\delta_3 - \theta_3)/\eta_3 + (1 - \theta_4/\delta_4)(1- \delta_3/\eta_3)\}(\gamma_1 - \eta_1)$
$+ \{(\delta_2 - \theta_2)/\gamma_2 + (\delta_3 - \theta_3)/\eta_3(1-\eta_2/\gamma_2) + (1-\theta_4/\delta_4)\{(\eta_2 - \delta_2)/\gamma_2 + (1- \delta_3/\eta_3)(1-\eta_2/\delta_2)\}\}(\phi_1 - \gamma_1)$

Linear models for Y, M and X

Consider the following models:
$E[Y|r,x,m,c] = \theta_0 + \theta_1 r + \theta_2 x + \theta_3 m + \theta_4' c$
$E[M|r,x,c] = \beta_0 + \beta_1 r + \beta_2 x + \beta_3' c$
$E[X|r,c] = \alpha_0 + \alpha_1 r + \alpha_2' c$

The results follow under these linear models for Proposition 4, to set the distribution of test scores M among black persons to its distribution among white persons, since:

$\mu_m = \Sigma_{x,m}\ E[Y|R=1,x,m,c]P(m|R=0,c)P(x|R=1,c)$
$= \Sigma_{x,m}\ (\theta_0 + \theta_1 + \theta_2 x + \theta_3 m + \theta_4' c)P(m|R=0,c)P(x|R=1,c)$
$= \theta_0 + \theta_1 + \theta_2 E[X|R=1,c] + \theta_3 E[M|R=0,c] + \theta_4' c$

We also have that:
$E[Y|R=1,c] = \Sigma_{x,m}\ E[Y|R=1,x,m,c]\ P(m|R=1,x,c)P(x|R=1,c).$
$= \Sigma_{x,m}\ (\theta_0 + \theta_1 + \theta_2 x + \theta_3 m + \theta_4' c)P(m|R=1,x,c)P(x|R=1,c).$
$= \theta_0 + \theta_1 + \theta_2\ E[X|R=1,c] + \theta_3 E[M|R=1,c] + \theta_4' c$
Thus, $E[Y|R=1,c] - \mu_m = \theta_3\ \{E[M|R=1,c] - E[M|R=0,c]\}$

Also:
$E[Y|R=0,c] = \Sigma_{x,m}\ E[Y|R=0,x,m,c]\ P(m|R=0,x,c)P(x|R=0,c).$
$= \Sigma_{x,m}\ (\theta_0 + \theta_2 x + \theta_3 m + \theta_4' c)P(m|R=0,x,c)P(x|R=0,c).$
$= \theta_0 + \theta_2\ E[X|R=0,c] + \theta_3 E[M|R=0,c] + \theta_4' c$
Thus, $\mu_m - E[Y|R=0,c] = \theta_1 + \theta_2\ \{E[X|R=1,c] - E[X|R=0,c]\}$

Note that:
$E[M|R=1,c] - E[M|R=0,c]$
$= \Sigma_x\ E[M|R=1,x,c]P(x|R=1,c) - \Sigma_x\ E[M|R=0,x,c]P(x|R=0,c)$
$= \Sigma_x\ (\beta_0 + \beta_1 + \beta_2 x + \beta_3' c)P(x|R=1,c) - \Sigma_x\ (\beta_0 + \beta_2 x + \beta_3' c)P(x|R=0,c)$
$= \beta_1 + \beta_2\ \{E[X|R=1,c] - E[X|R=0,c]\}$
Also:
$E[X|R=1,c] - E[X|R=0,c] = \alpha_1$

Thus, the remaining disparity is:
$\mu_m - E[Y|R=0,c] = \theta_1 + \theta_2\ \alpha_1$

And the disparity reduction is:
$E[Y|R=1,c] - \mu_m = \theta_3\ \{\beta_1 + \beta_2\ \alpha_1\}$

Successive logistic models for a rare binary outcome Y

Consider the following models:
Logit $P[Y|r,x,m,c] = \theta_0 + \theta_1 r + \theta_2 x + \theta_3 m + \theta_4'c$
Logit $P[Y|r,x,c] = \gamma_0 + \gamma_1 r + \gamma_2 x + \gamma_4'c$
Logit $P[Y|r,c] = \phi_0 + \phi_1 r + \phi_4'c$

The results under logistic models for Proposition 4, to set the distribution of test scores M among black persons to its distribution among white persons, follow since:

Under the assumption Logit $P[Y|\cdot] \approx \log P[Y|\cdot]$,
Logit μ_m
$\approx \text{Log}\{\Sigma_{x,m} P[Y|R=1,x,m,c] P(m|R=0,x,c)P(x|R=1,c)\}$
$= \text{Log}\{\Sigma_{x,m} \exp(\theta_0 + \theta_1 + \theta_2 x + \theta_3 m + \theta_4'c)P(m|R=0,x,c)P(x|R=1,c)\}$
$= \text{Log}\{\exp(\theta_0 + \theta_1 + \theta_4'c)\, E[\exp(\theta_2 X)|R=1,c]\, E[\exp(\theta_3 M)|R=0,c]\}$
$= \theta_0 + \theta_1 + \theta_4'c + \log E[\exp(\theta_2 X)|R=1,c] + \log E[\exp(\theta_3 M)|R=0,c]\}$

Similarly Logit $E[Y|R=1,c]$
$\approx \text{Log}\{\Sigma_{x,m} P[Y|R=1,x,m,c] P(m|R=1,x,c)P(x|R=1,c)\}$
$= \text{Log}\{\Sigma_{x,m} \exp(\theta_0 + \theta_1 + \theta_2 x + \theta_3 m + \theta_4'c)P(m|R=1,x,c)P(x|R=1,c)\}$
$= \text{Log}\{\exp(\theta_0 + \theta_1 + \theta_4'c)\, E[\exp(\theta_2 X)|R=1,c]\, E[\exp(\theta_3 M)|R=1,c]\}$
$= \theta_0 + \theta_1 + \theta_4'c + \log E[\exp(\theta_2 X)|R=1,c] + \log E[\exp(\theta_3 M)|R=1,c]$

Similarly Logit $E[Y|R=0,c]$
$\approx \text{Log}\{\Sigma_{x,m} P[Y|R=0,x,m,c] P(m|R=0,x,c)P(x|R=0,c)\}$
$= \text{Log}\{\Sigma_{x,m} \exp(\theta_0 + \theta_2 x + \theta_3 m + \theta_4'c)P(m|R=0,x,c)P(x|R=0,c)\}$
$= \text{Log}\{\exp(\theta_0 + \theta_4'c)\, E[\exp(\theta_2 X)|R=0,c]\, E[\exp(\theta_3 M)|R=0,c]\}$
$= \theta_0 + \theta_4'c + \log E[\exp(\theta_2 X)|R=0,c] + \log E[\exp(\theta_3 M)|R=0,c]$

Note that:
Logit $P[Y|R=1,c]$-Logit $P[Y|R=0,c]$
$\approx \text{Log}\, P[Y|R=1,c]$-Log $P[Y|R=0,c]$
$= \phi_1$

Also note that
Logit $P[Y|R=1,c]$-Logit $P[Y|R=0,c]$
$\approx \text{Log}\{\Sigma_x P[Y|R=1,x,c] P(x|R=1,c)\} - \text{Log}\{\Sigma_x P[Y|R=0,x,c] P(x|R=0,c)\}$
$= \text{Log}\{\Sigma_x \exp(\gamma_0 + \gamma_1 + \gamma_2 x + \gamma_4'c) P(x|R=1,c)\} - \text{Log}\{\Sigma_x \exp(\gamma_0 + \gamma_2 x + \gamma_4'c) P(x|R=0,c)\}$
$= \text{Log}\{\exp(\gamma_0 + \gamma_1 + \gamma_4'c)E[\exp(\gamma_2 X)|R=1,c]\} - \text{Log}\{\exp(\gamma_0 + \gamma_4'c)E[\exp(\gamma_2 X)|R=1,c]\}$
$= \gamma_1 + \text{Log}\, E[\exp(\gamma_2 X)|R=1,c] - \text{Log}\, E[\exp(\gamma_2 X)|R=0,c]$
$= \gamma_1 + \gamma_2 E[X|R=1,c] + \frac{1}{2}(\gamma_2)^2\sigma_X^2 - \gamma_2 E[X|R=0,c] - \frac{1}{2}(\gamma_2)^2\sigma_X^2$
$= \gamma_1 + \gamma_2\{E[X|R=1,c] - E[X|R=0,c]\}$
And so
$\phi_1 = \gamma_1 + \gamma_2\{E[X|R=1,c] - E[X|R=0,c]\}$
$\{E[X|R=1,c] - E[X|R=0,c]\} = (\phi_1 - \gamma_1)/\gamma_2$

Thus, the remaining disparity is equal to
Logit μ_m -Logit $E[Y|R=0]$
$= \theta_1 + \log E[\exp(\theta_2 X)|R=1,c]$-log $E[\exp(\theta_2 X)|R=0,c]$
$= \theta_1 + \log\{\exp(\theta_2 E[X|R=1,c] + \frac{1}{2}(\theta_2)^2\sigma_X^2)\} - \log\{\exp(\theta_2 E[X|R=0,c] + \frac{1}{2}(\theta_2)^2\sigma_X^2)\}$

$= \theta_1 + \theta_2 \{ E[X|R=1,c]-E[X|R=0,c]\}$
$= \theta_1 + \theta_2(\phi_1-\gamma_1)/\gamma_2$

And the disparity reduction is equal to
Logit $E[Y|R=1]$- Logit μ_m
= (Logit $E[Y|R=1]$- Logit $E[Y|R=0]$)-(Logit $E[Y|R=1]$- Logit μ_{xm})
$= \gamma_1 + \gamma_2 \{E[X|R=1,c] - E[X|R=0,c]\} - \theta_1 - \theta_2 \{ E[X|R=1,c]-E[X|R=0,c]\}$
$= (\gamma_1.\theta_1)+(1-\theta_2/\gamma_2)(\phi_1-\gamma_1)$

<u>Logistic model for a rare binary outcome Y with linear models for M and X</u>

Consider the following models:
Logit $P[Y|r,x,m,c] = \theta_0 + \theta_1 r + \theta_2 x + \theta_3 m + \theta_4' c$
$E[M|r,x,c] = \beta_0 + \beta_1 r + \beta_2 x + \beta_4' c$
$E[X|r,c] = \alpha_0 + \alpha_1 r + \alpha_4' c$

Assume the outcome is rare and the error term in the model for $E[X|r,c]$ is normally distributed and constant variance σ_x, and the error term in the model for $E[M|r,x,c]$ is normally distributed with constant variance σ_M

The results under these models for Proposition 4, to set the distribution of test scores M among black persons to its distribution among white persons, follow since:

Under the assumption Logit $P[Y|\cdot] \approx \log P[Y|\cdot]$, we have that Logit μ_{xm}
$\approx$ Log $\{\Sigma_{x,m} P[Y|R=1,x,m,c] P(m|R=0,x,c)P(x|R=1,c)\}$
= Log $\{\Sigma_{x,m} \exp(\theta_0 + \theta_1 + \theta_2 x + \theta_3 m + \theta_4' c)P(m|R=0,x,c)P(x|R=1,c)\}$
= Log $\{\exp(\theta_0 + \theta_1 + \theta_4' c)\ E[\exp(\theta_2 X)|R=1,c]\ E[\exp(\theta_3 M)|R=0,c]\}$
= Log $\{\exp(\theta_0 + \theta_1 + \theta_4' c)\ \exp((\theta_2)(\alpha_0 + \alpha_1 + \alpha_4' c) + \frac{1}{2}(\theta_2)^2 \sigma_x^2)\ E[\exp(\theta_3 M)|R=0,c]\}$
= Log $\{\exp(\theta_0 + \theta_1 + \theta_4' c)\ \exp((\theta_2)(\alpha_0 + \alpha_1 + \alpha_4' c) + \frac{1}{2}(\theta_2)^2 \sigma_x^2)\ \exp((\theta_3)E[M|R=0,c] + \frac{1}{2}(\theta_3)^2 \sigma_M^2)\}$
$= \theta_0 + \theta_1 + \theta_4' c + \theta_2(\alpha_0 + \alpha_1 + \alpha_4' c) + \frac{1}{2}(\theta_2)^2 \sigma_x^2 + (\theta_3)E[M|R=0,c] + \frac{1}{2}(\theta_2)^2 \sigma_M^2$

Similarly Logit $E[Y|R=1,c]$
$\approx$ Log $\{\Sigma_{x,m} P[Y|R=1,x,m,c] P(m|R=1,x,c)P(x|R=1,c)\}$
= Log $\{\Sigma_{x,m} \exp(\theta_0 + \theta_1 + \theta_2 x + \theta_3 m + \theta_4' c)P(m|R=1,x,c)P(x|R=1,c)\}$
= Log $\{\exp(\theta_0 + \theta_1 + \theta_4' c)\ E[\exp(\theta_2 X)|R=1,c]\ E[\exp(\theta_3 M)|R=1,c]\}$
= Log $\{\exp(\theta_0 + \theta_1 + \theta_4' c)\ \exp((\theta_2)(\alpha_0 + \alpha_1 + \alpha_4' c) + \frac{1}{2}(\theta_2)^2 \sigma_x^2)\ \exp((\theta_3)E[M|R=1,c] + \frac{1}{2}(\theta_3)^2 \sigma_M^2)\}$
$= \theta_0 + \theta_1 + \theta_4' c + \theta_2(\alpha_0 + \alpha_1 + \alpha_4' c) + \frac{1}{2}(\theta_2)^2 \sigma_x^2 + (\theta_3)E[M|R=1,c] + \frac{1}{2}(\theta_2)^2 \sigma_M^2$

Similarly Logit $E[Y|R=0,c]$
$\approx$ Log $\{\Sigma_{x,m} P[Y|R=0,x,m,c] P(m|R=0,x,c)P(x|R=0,c)\}$
= Log $\{\Sigma_{x,m} \exp(\theta_0 + \theta_2 x + \theta_3 m + \theta_4' c)P(m|R=0,x,c)P(x|R=0,c)\}$
= Log $\{\exp(\theta_0 + \theta_4' c)\ E[\exp(\theta_2 X)|R=0,c]\ E[\exp(\theta_3 M)|R=0,c]\}$
= Log $\{\exp(\theta_0 + \theta_4' c)\ \exp((\theta_2)(\alpha_0 + \alpha_4' c) + \frac{1}{2}(\theta_2)^2 \sigma_x^2)\ \exp((\theta_3)E[M|R=0,c] + \frac{1}{2}(\theta_3)^2 \sigma_M^2)\}$
$= \theta_0 + \theta_4' c + \theta_2(\alpha_0 + \alpha_4' c) + \frac{1}{2}(\theta_2)^2 \sigma_x^2 + (\theta_3)E[M|R=0,c] + \frac{1}{2}(\theta_2)^2 \sigma_M^2$

Note that logit $E[M|R=1,c]$ – logit $E[M|R=0,c] \approx$
$= \Sigma_x E[M|R=1,x,c]P(x|R=1,c) - \Sigma_x E[M|R=0,x,c]P(x|R=0,c)$

$= \Sigma_x (\beta_0 + \beta_1 + \beta_2 x + \beta_3'c)P(x|R=1,c) - \Sigma_x (\beta_0 + \beta_2 x + \beta_3'c)P(x|R=0,c)$
$= \beta_1 + \beta_2 \{E[X|R=1,c] - E[X|R=0,c]\}$
$= \beta_1 + \beta_2 \alpha_1$

Thus the disparity reduction $E[Y|R=1,c]/\mu_{xm}$ is
$= \exp(\text{Logit } E[Y|R=1,c] - \text{Logit } \mu_{xm})$
$\approx \exp(\text{Log } E[Y|R=1,c] - \text{Log } \mu_{xm})$
$= \exp(\theta_3\{E[M|R=1,c]-E[M|R=0,c]\})$
$= \exp(\theta_3\{\beta_1 + \beta_2\alpha_1\})$

And the remaining disparity $\mu_{xm}/E[Y|R=0,c]$ is
$= \exp(\text{Logit } \mu_{xm} - \text{Logit } E[Y|R=0,c])$
$\approx \exp(\text{Log } \mu_{xm} - \text{Log } E[Y|R=0,c])$
$= \exp(\theta_1 + \theta_2\alpha_1)$

<u>Oaxaca-Blinder decomposition</u>

Consider the two sets of race-stratified linear models that each can be used to carry out different Oaxaca-Blinder decompositions:

Set 1:
$E[Y|R=1,x,c]=\omega_0 + \omega_1 x + \omega_3'c$
$E[Y|R=0,x,c]=\pi_0 + \pi_1 x + \pi_3'c$

To simplify the formulas we derive, we assume that $\omega_3=\pi_3$. We could allow for $\omega_3 \neq \pi_3$ but this is does not materially affect our proof that propositions 1-4 can be expressed as causal implementations of the Oaxaca-Blinder decomposition.

Set 2:
$E[Y|R=1,m,x,c]=\alpha_0 + \alpha_1 x + \alpha_2 m + \alpha_3'c$
$E[Y|R=0,m,x,c]=\beta_0 + \beta_1 x + \beta_2 m + \beta_3'c$

Consider also successive linear models for Y, this time with interaction terms between R and X and also R and M. (These models could allow for interactions between R and C, and while this would slightly change some of the formulas we derive, this additional complexity does not affect the ability to express propositions 1-4 as causal implementations of the Oaxaca-Blinder decomposition).

Set 3:
$E[Y|r,x,m,c] = \theta_0 + \theta_1 r + \theta_2 x + \theta_3 m + \theta_4 rx + \theta_5 rm + \theta_6'c$
$E[Y|r,x,c] = \gamma_0 + \gamma_1 r + \gamma_2 x + \gamma_4 rx + \gamma_6'c$
$E[Y|r,c] = \phi_0 + \phi_1 r + \phi_6'c$

Again, we could incorporate interaction terms between race and the covariates C in these models, but again, this additional complexity would not affect the ability to express propositions 1-4 as causal implementations of the Oaxaca-Blinder decomposition.

For Proposition 1 (i.e. equalize the distribution of childhood SES X across race R), the results under an Oaxaca-Blinder decomposition with models from set 1 equate to results using linear models from set 3 since, under assumption A1:

$\mu_x = \Sigma_x E[Y|R=1,x,c]P(x|R=0,c)$
$= \Sigma_x (\gamma_0 + \gamma_1 + \gamma_2 x + \gamma_4 x + \gamma_6'c)P(x|R=0,c)$
$= \gamma_0 + \gamma_1 + (\gamma_2 + \gamma_4) E[X|R=0,c] + \gamma_6'c$
Similarly,
$E[Y|R=0,c] = E[Y|R=0,x,c]P(x|R=0,c)$
$= \Sigma_x (\gamma_0 + \gamma_2 x + \gamma_6'c)P(x|R=0,c)$
$= \gamma_0 + \gamma_2 E[X|R=0,c] + \theta_6'c$
Thus, $\mu_x - E[Y|R=0,c] = \gamma_1 + \gamma_4 E[X|R=0,c]$
Also,
$E[Y|R=1,c] = E[Y|R=1,x,m,c]P(x|R=1,c)$
$= \Sigma_x (\gamma_0 + \gamma_1 r + \gamma_2 x + \gamma_4 x + \gamma_6'c)P(x|R=0,c)$
$= (\gamma_0 + \gamma_1 + (\gamma_2 + \gamma_4) E[X|R=1,c] + \gamma_6'c$
Thus, $E[Y|R=1,c] - \mu_x = (\gamma_2 + \gamma_4) \{E[X|R=1,c] - E[X|R=0,c]\}$

Note that
$E[Y|R=1,c] = \Sigma_x E[Y|R=1,x,c]P(x|R=1,c)$
$= \Sigma_x (\omega_0 + \omega_1 x + \omega_3'c) P(x|R=1,c)$
$= \omega_0 + \omega_1 E[X|R=1,c] + \omega_3'c$
Similarly,
$E[Y|R=0,c] = \Sigma_x E[Y|R=0,x,c]P(x|R=0,c)$
$= \Sigma_x (\pi_0 + \pi_1 x + \pi_3'c)P(x|R=0,c)$
$= \pi_0 + \pi_1 E[X|R=0,c] + \pi_3'c$

Thus,
$E[Y|R=1,c] - E[Y|R=0,c]$
$= (\omega_0 - \pi_0) + \omega_1 E[X|R=1,c] - \pi_1 E[X|R=0,c]$
$= (\omega_0 - \pi_0) + \omega_1 E[X|R=1,c] - \pi_1 E[X|R=0,c] + \omega_1 E[X|R=0,c] - \omega_1 E[X|R=0,c]$
$= (\omega_0 - \pi_0) + (\omega_1 - \pi_1) E[X|R=0,c] + \omega_1 \{E[X|R=1,c] - E[X|R=0,c]\}$

In an Oaxaca-Blinder decomposition, the terms $(\omega_0 - \pi_0)$ and $(\omega_1 - \pi_1)E[X|R=0,c]$ could be referred to as the "unexplained portion" and the term $\omega_1\{E[X|R=1,c] - E[X|R=0,c]\}$ could be referred to as the "explained" portion, whose sum equals the total disparity $E[Y|R=1,c] - E[Y|R=0,c]$.

Note that by definition, $(\omega_0 - \pi_0) = \gamma_1$, $\pi_1 = \gamma_2$, and $(\omega_1 - \pi_1) = \gamma_4$.

Thus,
$\mu_x - E[Y|R=0,c]$
$= \gamma_1 + \gamma_4 E[X|R=0,c]$
$= (\omega_0 - \pi_0) + (\omega_1 - \pi_1) E[X|R=0,c]$
Also,
$E[Y|R=1,c] - \mu_x$
$= (\gamma_2 + \gamma_4)\{E[X|R=1,c] - E[X|R=0,c]\}$
$= \omega_1\{E[X|R=1,c] - E[X|R=0,c]\}$

Thus, under assumption A1, these quantities can be interpreted as the residual disparity and disparity reduction under an intervention to equalize X alone (proposition 1). Note that if $\gamma_4=0$ such that $\omega_1=\pi_1$ we obtain the results under linear models in the main text.

For Proposition 2 (i.e. equalize the distribution of test scores M across race R within levels of childhood SES X), the results under an Oaxaca-Blinder decomposition with models from set 2 equate to results using linear models from set 3 since under assumption A2:

$\mu_{m|x} = \Sigma_m E[Y|R=1,x,m,c]P(m|R=0,x,c)$
$= \Sigma_x (\theta_0 + \theta_1 + \theta_2 x + \theta_3 m + \theta_4 x + \theta_5 m + \theta_6'c)P(m|R=0,x,c)$
$= \theta_0 + \theta_1 + (\theta_2 + \theta_4)\, x + (\theta_3 + \theta_5)\, E[M|R=0,x,c] + \theta_6'c$
Similarly,
$E[Y|R=0,x,c] = \Sigma_m E[Y|R=0,x,m,c]P(m|R=0,x,c)$
$= \Sigma_m (\theta_0 + \theta_2 x + \theta_3 m + \theta_6'c)P(m|R=0,x,c)$
$= \theta_0 + \theta_2 x + \theta_5 E[M|R=0,x,c] + \theta_6'c$
Thus, $\mu_{m|x}$ -$E[Y|R=0,c] = \theta_1 + \theta_4 x + \theta_5 E[M|R=0,x,c]$
Also,
$E[Y|R=1,x,c] = E[Y|R=1,x,m,c]P(x|R=1,c)$
$= \Sigma_m (\theta_0 + \theta_1 + \theta_2 x + \theta_3 m + \theta_4 x + \theta_5 m + \theta_6'c)P(m|R=1,x,c)$
$= \theta_0 + \theta_1 + (\theta_2 + \theta_4)\, x + (\theta_3 + \theta_5)\, E[M|R=1,x,c] + \theta_6'c$
Thus, $E[Y|R=1,x,c]- \mu_{m|x} = (\theta_3 + \theta_5)\{E[M|R=1,x,c] - E[M|R=0,x,c]\}$

Note that
$E[Y|R=1,x,c]= \Sigma_m E[Y|R=1,m,x,c]P(m|R=1,x,c)$
$=\Sigma_m (\alpha_0 + \alpha_1 x + \alpha_2 m + \alpha_3'c)\, P(m|R=1,x,c)$
$= \alpha_0 + \alpha_1 x + \alpha_2 E[M|R=1,x,c] + \alpha_3'c$
Similarly,
$E[Y|R=0,x,c]= \Sigma_m E[Y|R=0,m,x,c]P(m|R=0,x,c)$
$=\Sigma_m (\beta_0 + \beta_1 x + \beta_2 m + \beta_3'c)P(m|R=0,x,c)$
$= \beta_0 + \beta_1 x + \beta_2 E[M|R=0,x,c] + \beta_3'c$
Thus,
$E[Y|R=1,x,c]-E[Y|R=0,x,c]$
$= (\alpha_0-\beta_0) + \alpha_1 x - \beta_1 x + \alpha_2 E[M|R=1,x,c]- \beta_2 E[M|R=0,x,c]$
$= (\alpha_0-\beta_0) + \alpha_1 x - \beta_1 x + \alpha_2 E[M|R=1,x,c]- \beta_2 E[M|R=0,x,c] + \alpha_2 E[M|R=0,x,c] - \alpha_2 E[M|R=0,x,c]$
$= (\alpha_0-\beta_0) + (\alpha_1-\beta_1)x + (\alpha_2-\beta_2)E[M|R=0,x,c] + \alpha_2\{E[M|R=1,x,c]-E[X|R=0,x,c]\}$

In an Oaxaca-Blinder decomposition, the terms $(\alpha_0-\beta_0)$ and $(\alpha_1-\beta_1)x$ and $(\alpha_2-\beta_2)E[M|R=0,x,c]$ could be referred to as the "unexplained portion given X." The term $\alpha_2\{E[M|R=1,x,c]-E[X|R=0,x,c]\}$ could be referred to as the "explained portion given X," whose sum equals the total disparity within levels of X i.e. $E[Y|R=1,x,c]-E[Y|R=0,x,c]$.

Note that by definition $(\alpha_0-\beta_0)=\theta_1$, $(\alpha_1-\beta_1)=\theta_4$, $\beta_2=\theta_3$, and $(\alpha_2-\beta_2) = \theta_5$.

Thus,
$\mu_{m|x}$ $-E[Y|R=0,c]$
$= \theta_1 + \theta_4 x + \theta_5 E[M|R=0,x,c]$
$= (\alpha_0-\beta_0) + (\alpha_1-\beta_1)x + (\alpha_2-\beta_2)E[M|R=0,x,c]$
Also
$E[Y|R=1,x,c]- \mu_{m|x}$

$= (\theta_3 + \theta_5)\{E[M|R=1,x,c] - E[M|R=0,x,c]\}$
$= \alpha_2\{E[M|R=1,x,c] - E[M|R=0,x,c]\}$

Thus, under assumption A2, these quantities can be interpreted as the residual disparity and disparity reduction under an intervention to equalize M within levels of X (proposition 2). Note that if $\theta_4=0$ such that $\alpha_1=\beta_1$ and $\theta_5=0$ such that $\alpha_2=\beta_2$ we obtain the results under linear models in the main text.

For Proposition 3 (i.e. equalize the distribution of childhood SES X and test scores M across race R), the results under an Oaxaca-Blinder decomposition with models from set 2 equate to results using linear models from set 3 since under assumptions A1' and A2':

$\mu_{xm} = \Sigma_{xm}\, E[Y|R=1,m,x,c]\, P(m|R=0,x,c)P(x|R=0,c)$
$= \Sigma_{xm}\, (\theta_0 + \theta_1 + \theta_2 x + \theta_3 m + \theta_4 x + \theta_5 m + \theta_6'c)\, P(m|R=0,x,c)P(x|R=0,c)$
$= \theta_0 + \theta_1 + (\theta_2 + \theta_4)\, E[X|R=0,c] + (\theta_3 + \theta_5)\, E[M|R=0,c] + \theta_6'c$
Similarly,
$E[Y|R=0,c] = \Sigma_{xm}\, E[Y|R=0,m,x,c]\, P(m|R=0,x,c)P(x|R=0,c)$
$= \Sigma_{xm}\, (\theta_0 + \theta_2 x + \theta_3 m + \theta_6'c)\, P(m|R=0,x,c)P(x|R=0,c)$
$= \theta_0 + \theta_2 E[X|R=0,c] + \theta_3 E[M|R=0,c] + \theta_6'c$
Thus, $\mu_x - E[Y|R=0,c] = \theta_1 + \theta_4 E[X|R=0,c] + \theta_5 E[M|R=0,c]$
Also,
$E[Y|R=1,c] = E[Y|R=1,x,m,c]P(x|R=1,c)$
$= \Sigma_{xm}\, (\theta_0 + \theta_1 + \theta_2 x + \theta_3 m + \theta_4 x + \theta_5 m + \theta_6'c)\, P(m|R=1,x,c)P(x|R=1,c)$
$= \theta_0 + \theta_1 + (\theta_2 + \theta_4)\, E[X|R=1,c] + (\theta_3 + \theta_5)\, E[M|R=1,c] + \theta_6'c$
Thus, $E[Y|R=1,c] - \mu_{xm} = (\theta_2 + \theta_4)\, \{E[X|R=1,c] - E[X|R=0,c]\} + (\theta_3 + \theta_5)\, \{E[M|R=1,c] - E[M|R=0,c]\}$

Note that
$E[Y|R=1,c] = \Sigma_{xm}\, E[Y|R=1,m,x,c]\, P(m|R=1,x,c)P(x|R=1,c)$
$= \Sigma_{xm}\, (\alpha_0 + \alpha_1 x + \alpha_2 m + \alpha_3'c)\, P(m|R=1,x,c)P(x|R=1,c)$
$= \alpha_0 + \alpha_1 E[X|R=1,c] + \alpha_2 E[M|R=1,c] + \alpha_3'c$
Similarly we have that:
$E[Y|R=0,c] = \Sigma_{xm}\, E[Y|R=0,x,c]\, P(m|R=0,x,c)P(x|R=0,c)$
$= \Sigma_{xm}\, (\beta_0 + \beta_1 x + \beta_2 m + \beta_3'c)\, P(m|R=0,x,c)P(x|R=0,c)$
$= \beta_0 + \beta_1 E[X|R=0,c] + \beta_2 E[M|R=0,c] + \beta_3'c$
Thus,
$E[Y|R=1,c] - E[Y|R=0,c]$
$= (\alpha_0 - \beta_0) + \alpha_1 E[X|R=1,c] - \beta_1 E[X|R=0,c] + \alpha_2 E[M|R=1,c] - \beta_2 E[M|R=0,c]$
$= (\alpha_0 - \beta_0) + \alpha_1 E[X|R=1,c] - \beta_1 E[X|R=0,c] + \alpha_2 E[M|R=1,c] - \beta_2 E[M|R=0,c]$
$+ \alpha_1 E[X|R=0,c] - \alpha_1 E[X|R=0,c] + \alpha_2 E[M|R=0,c] - \alpha_2 E[M|R=0,c]$
$= (\alpha_0 - \beta_0) + (\alpha_1 - \beta_1) E[X|R=0,c] + (\alpha_2 - \beta_2) E[M|R=0,c] + \alpha_1\{E[X|R=1,c] - E[X|R=0,c]\} + \alpha_2\{E[M|R=1,c] - E[M|R=0,c]\}$

In an Oaxaca-Blinder decomposition, the terms $(\alpha_0 - \beta_0)$ and $(\alpha_1 - \beta_1)E[X|R=0,c]$ and $(\alpha_2 - \beta_2)E[M|R=0,c]$ would be referred to as the "unexplained portion" and the third term $\alpha_1\{E[X|R=1,c] - E[X|R=0,c]\}$ and fourth term $\alpha_2\{E[M|R=1,c] - E[M|R=0,c]\}$ would be referred to as the "explained" portion.

Note that by definition, $(\alpha_0 - \beta_0) = \theta_1$, $\beta_1 = \theta_2$, $\beta_2 = \theta_3$, $(\alpha_1 - \beta_1) = \theta_4$, and $(\alpha_2 - \beta_2) = \theta_5$.

Thus,
μ_x -E[Y|R=0,c]
$= \theta_1 + \theta_4 E[X|R=0,c] + \theta_5 E[M|R=0,c]$
$= (\alpha_0-\beta_0) + (\alpha_1-\beta_1)E[X|R=0,c] + (\alpha_2-\beta_2)E[M|R=0,c]$
Also,
$E[Y|R=1,c]- \mu_x$
$= (\theta_2 + \theta_4) \{E[X|R=1,c] - E[X|R=0,c]\} + (\theta_3 + \theta_5) \{E[M|R=1,c] - E[M|R=0,c]\}$
$= \alpha_1\{E[X|R=1,c] - E[X|R=0,c]\} + \alpha_2\{E[M|R=1,c] - E[M|R=0,c]\}$

Thus, under assumptions A1' and A2', these quantities can be interpreted as the residual disparity and disparity reduction under an intervention to equalize X and M (proposition 3). Note that if $\theta_4=0$ such that $\alpha_1=\beta_1$ and $\theta_5=0$ such that $\alpha_2=\beta_2$ we obtain the results under linear models in the main text.

For Proposition 4 (i.e. equalize the distribution of test scores M across race R), the results under a detailed Oaxaca-Blinder decomposition, with models from set 2, can be used to obtain results using linear models from set 3, since under assumption A2:

$\mu_m = \Sigma_{xm} E[Y|R=1,m,x,c] P(m|R=0,x,c)P(x|R=1,c)$
$= \Sigma_{xm} (\theta_0 + \theta_1 + \theta_2 x + \theta_3 m + \theta_4 x + \theta_5 m + \theta_6'c) P(m|R=0,c)P(x|R=1,c)$
$= \theta_0 + \theta_1 + (\theta_2 + \theta_4) E[X|R=1,c] + (\theta_3 + \theta_5) E[M|R=0,c] + \theta_6'c$
Similarly,
$E[Y|R=0,c] = \Sigma_{xm} E[Y|R=0,m,x,c] P(m|R=0,x,c)P(x|R=0,c)$
$= \Sigma_{xm} (\theta_0 + \theta_2 x + \theta_3 m + \theta_6'c) P(m|R=0,x,c)P(x|R=0,c)$
$= \theta_0 + \theta_2 E[X|R=0,c] + \theta_3 E[M|R=0,c] + \theta_6'c$
Thus, μ_x -E[Y|R=0,c] $= \theta_1 + \theta_2\{E[X|R=1,c]-E[X|R=0,c]\} + \theta_4 E[X|R=1,c]+ \theta_5 E[M|R=0,c]$
Also,
$E[Y|R=1,c] = E[Y|R=1,x,m,c]P(x|R=1,c)$
$= \Sigma_{xm} (\theta_0 + \theta_1 + \theta_2 x + \theta_3 m + \theta_4 x + \theta_5 m + \theta_6'c) P(m|R=1,x,c)P(x|R=1,c)$
$= \theta_0 + \theta_1 + (\theta_2 + \theta_4) E[X|R=1,c] + (\theta_3 + \theta_5) E[M|R=1,c] + \theta_6'c$
Thus, $E[Y|R=1,c]- \mu_m = (\theta_3 + \theta_5) \{E[M|R=1,c] - E[M|R=0,c]\}$

Note that
$E[Y|R=1,c]= \Sigma_{xm} E[Y|R=1,m,x,c] P(m|R=1,x,c)P(x|R=1,c)$
$=\Sigma_{xm} (\alpha_0 + \alpha_1 x + \alpha_2 m + \alpha_3'c) P(m|R=1,x,c)P(x|R=1,c)$
$= \alpha_0 + \alpha_1 E[X|R=1,c] + \alpha_2 E[M|R=1,c] + \alpha_3'c$
Similarly we have that:
$E[Y|R=0,c]= \Sigma_x E[Y|R=0,x,c] P(m|R=0,x,c)P(x|R=0,c)$
$=\Sigma_x (\beta_0 + \beta_1 x + \beta_2 m + \beta_3'c) P(m|R=0,x,c)P(x|R=0,c)$
$= \beta_0 + \beta_1 E[X|R=0,c] + \beta_2 E[M|R=0,c] + \beta_3'c$
Thus,
$E[Y|R=1,c]-E[Y|R=0,c]$
$= (\alpha_0-\beta_0) + \alpha_1 E[X|R=1,c]- \beta_1 E[X|R=0,c] + \alpha_2 E[M|R=1,c]- \beta_2 E[M|R=0,c]$
$= (\alpha_0-\beta_0) + \alpha_1 E[X|R=1,c]- \beta_1 E[X|R=0,c] + \alpha_2 E[M|R=1,c]- \beta_2 E[M|R=0,c]$
$+ \alpha_1 E[X|R=0,c] – \alpha_1 E[X|R=0,c] + \alpha_2 E[M|R=0,c] – \alpha_2 E[M|R=0,c]$
$= (\alpha_0-\beta_0) + (\alpha_1-\beta_1)E[X|R=0,c] + (\alpha_2-\beta_2)E[M|R=0,c] + \alpha_1\{E[X|R=1,c]-E[X|R=0,c]\} + \alpha_2\{E[M|R=1,c]-E[M|R=0,c]\}$

A so-called detailed Oaxaca-Blinder decomposition would refer to the terms $(\alpha_0-\beta_0)$ and $(\alpha_1-\beta_1)E[X|R=0,c]$ and $(\alpha_2-\beta_2)E[M|R=0,c]$ as the unexplained portion, and then partition the "explained" portion into the part independently explained by X i.e. $\alpha_1\{E[X|R=1,c]-E[X|R=0,c]\}$ and the part independently explained by M i.e. $\alpha_2\{E[M|R=1,c]-E[M|R=0,c]\}$, with all terms summing to equal the total disparity $E[Y|R=1,c]-E[Y|R=0,c]$.

Note that by definition, $(\alpha_0-\beta_0)=\theta_1$, $\beta_1=\theta_2$, $\beta_2=\theta_3$, $(\alpha_1-\beta_1)=\theta_4$, and $(\alpha_2-\beta_2)=\theta_5$.

Note also that
$(\alpha_1-\beta_1)E[X|R=0,c] + \alpha_1\{E[X|R=1,c]-E[X|R=0,c]\}$
$= \alpha_1E[X|R=1,c] - \beta_1E[X|R=0,c] + \beta_1E[X|R=1,c] - \beta_1E[X|R=1,c]$
$= (\alpha_1-\beta_1)E[X|R=1,c] + \beta_1\{E[X|R=1,c]-E[X|R=0,c]\}$

Thus, $\mu_m - E[Y|R=0,c]$
$= \theta_1 + \theta_2\{E[X|R=1,c]-E[X|R=0,c]\} + \theta_4E[X|R=1,c]+ \theta_5E[M|R=0,c]$
$= (\alpha_0-\beta_0) + \beta_1\{E[X|R=1,c]-E[X|R=0,c]\} + (\alpha_1-\beta_1)E[X|R=1,c]+ (\alpha_2-\beta_2)E[M|R=0,c]$
$= (\alpha_0-\beta_0) + \alpha_1\{E[X|R=1,c]-E[X|R=0,c]\} + (\alpha_1-\beta_1)E[X|R=0,c]+ (\alpha_2-\beta_2)E[M|R=0,c]$
Also,
$E[Y|R=1,c]- \mu_m$
$= (\theta_3 + \theta_5) \{E[M|R=1,c] - E[M|R=0,c]\}$
$= \alpha_2\{E[M|R=1,c] - E[M|R=0,c]\}$

Thus, under assumption A2, the residual disparity under an intervention to equalize M alone is in fact equal to the sum of the "unexplained" portion and the portion "independently explained" by X. The disparity reduction is equal to the portion "independently explained" by M. (Note that these formulae equate to the ones in the main text if $\theta_4=0$ such that $\alpha_1=\beta_1$ and $\theta_5=0$ such that $\alpha_2= \beta_2$).

This result provides some further intuition for why the disparity reduction under Proposition 4 does not generally equal the difference between reductions under Proposition 1 (equalize X alone) and Proposition 3 (equalize X and M). This would only be so under the special case $\alpha_1=\omega_1$ i.e. M does not mediate the effect of X (and that both assumptions A1 and A2 hold). Only in that special case could the portion "independently explained" by X be interpreted as the disparity reduction under an intervention to equalize X alone (i.e. Proposition 1).

References

1. Oaxaca R. Male-Female Wage Differentials in Urban Labor Markets. *Int Econ Rev (Philadelphia)*. 1973;14(3):693. doi:10.2307/2525981.
2. Blinder AS. Wage Discrimination: Reduced Form and Structural Estimates. *J Hum Resour*. 1973;8(4):436. doi:10.2307/144855.
3. Yun MS. Decomposing differences in the first moment. *Econ Lett*. 2004;82(2):275-280. doi:10.1016/j.econlet.2003.09.008.
4. Fairlie RW. An extension of the Blinder-Oaxaca decomposition technique to logit and probit models. *J Econ Soc Meas*. 2005;30(873):305-316. doi:http://iospress.metapress.com/content/0747-9662/.
5. Van Kerm P, Yu S, Choe C. Decomposing quantile wage gaps: a conditional likelihood approach. *J R Stat Soc Ser C Appl Stat*. 2016;65(4):507-527. doi:10.1111/rssc.12137.
6. Rothe C. Decomposing the Composition Effect: The Role of Covariates in Determining Between-Group Differences in Economic Outcomes. *J Bus Econ Stat*. 2015;33(3):323-337. doi:10.1080/07350015.2014.948959.
7. DiNardo J, Fortin NM, Lemieux T. Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. *Econometrica*. 1996;64(5):1001-1044.
8. Barsky R, Bound J, Charles KK, Lupton JP. Accounting for the black-white wealth gap: A nonparametric approach. *J Am Stat Assoc*. 2002;97(459):663-673. doi:10.1198/016214502388618401.
9. Kline P. Regression, Reweighting, or Both: Oaxaca-Blinder as a Reweighting Estimator. *Am Econ Rev*. 2011;101(3):532-537.
10. Elder TE, Godderis JH, Haider SJ. Isolating the roles of individual covariates in reweighting estimation. *J Appl Econom*. 2015;30(30):1169-1191. doi:10.1002/jae.2433.
11. Słoczyński T. The Oaxaca-Blinder Unexplained Component as a Treatment Effects Estimator. *Oxf Bull Econ Stat*. 2015;77(4):588-604. doi:10.1111/obes.12075.
12. Black D, Haviland A, Sanders S, Taylor L. Why Do Minority Men Earn Less ? A Study of Wage Differentials among the Highly. *Rev Econ Stat*. 2006;88(2):300-313.
13. Firpo S, Fortin NM, Lemieux T. Decomposing wage distributions using recentered influence function regressions. *mimeo, Univ Br Columbia ….* 2007:1-60.
14. Fortin N, Lemieux T, Firpo S. Decomposition Methods in Economics. In: *Handbook of Labor Economics*. Vol 4. ; 2011:1-102. doi:10.1016/S0169-7218(11)00407-2.
15. Huber M. Causal Pitfalls in the Decomposition of Wage Gaps. *J Bus Econ Stat*. 2015;33(2):179-191. doi:10.1080/07350015.2014.937437.
16. Kaufman JS. Epidemiologic analysis of racial/ethnic disparities: Some fundamental issues and a cautionary example. *Soc Sci Med*. 2008;66(8):1659-1669. doi:10.1016/j.socscimed.2007.11.046.
17. Greiner DJ, Rubin DB. Causal Effects of Perceived Immutable Characteristics. *Rev Econ Stat*. 2011;93(3):775-785. doi:10.1162/REST_a_00110.
18. Holland PW. Statistics and Causal Inference Authors (s): Paul W . Holland Source : Journal of the American Statistical Association , Vol . 81 , No . 396 (Dec ., 1986), pp . Published by : Taylor & Francis , Ltd . on behalf of the American Statistical Association . 2016;81(396):945-960.
19. Hernán MA, Robins JM. *Causal Inference*. (forthcoming): CRC Press; 2018.
20. Robins JM, Richardson TS. Alternative graphical causal models and the identification of direct effects. In: Shrout PE, Keyes KM, Ornstein K, eds. *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*. New York, NY: Oxford University Press, Inc.; 2011:103-158.
21. Pearl J. *Causality: Models, Reasoning and Inference*. 2nd ed. New York, NY: Cambridge University Press; 2009.

22. Reskin B. The Race Discrimination System. *Annu Rev Sociol*. 2012;38(1):17-35. doi:10.1146/annurev-soc-071811-145508.
23. Rothstein R. *The Color of Law*. New York, NY: Liveright Publishing Corporation; 2017.
24. Bailey ZD, Krieger N, Agénor M, Graves J, Linos N, Bassett MT. Structural racism and health inequities in the USA: evidence and interventions. *Lancet*. 2017;389(10077):1453-1463. doi:10.1016/S0140-6736(17)30569-X.
25. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992;3(2):143-155.
26. VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions, and composition. *Stat its Inference*. 2009;2:457-468.
27. Holland PW. Statistics and Causal Inference. *J Am Stat Assoc*. 1986;81(396):945-960. doi:10.1080/01621459.1986.10478354.
28. Avin C, Shpitser I, Pearl J. Identifiability of Path-Specific Effects. In: *Proceedings of International Joint Conference on Artificial Intelligence*. Edinburgh, Scotland; 2005:357-363.
29. Albert JM, Nelson S. Generalized Causal Mediation Analysis. *Biometrics*. 2011;67(3):1028-1038. doi:10.1111/j.1541-0420.2010.01547.x.
30. VanderWeele T, Vansteelandt S. Mediation Analysis with Multiple Mediators. *Epidemiol Method*. 2014;2(1):95-115. doi:10.1515/em-2012-0010.
31. VanderWeele TJ, Vansteelandt S, Robins JM. Effect Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder. *Epidemiology*. 2014;25(2):300-306. doi:10.1097/EDE.0000000000000034.
32. Steen J, Loeys T, Moerkerke B, Vansteelandt S. Flexible Mediation Analysis With Multiple Mediators. *Am J Epidemiol*. 2017;186(2):184-193. doi:10.1093/aje/kwx051.
33. Robins JM. Semantics of causal DAG models and the identification of direct and indirect effects. In: *Highly Structured Stochastic Systems*. ; 2003:70-81.
34. Imai K, Yamamoto T. Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Polit Anal*. 2013;21(2):141-171. doi:10.1093/pan/mps040.
35. Daniel RM, De Stavola BL, Cousens SN, Vansteelandt S. Causal mediation analysis with multiple mediators. *Biometrics*. 2015;71(1):1-14. doi:10.1111/biom.12248.
36. Didelez V, Dawid AP, Geneletti S. Direct and Indirect Effects of Sequential Treatments. In: Detcher R, Richardson T, eds. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. AUIA Press; 2006:138-146.
37. Geneletti S. Identifying Direct and Indirect Effects in a Non-Counterfactual Framework. *J R Stat Soc Ser B*. 2007;69(2):199-215.
38. Zheng W, van der Laan MJ. Causal Mediation in a Survival Setting with Time-Dependent Mediators. *UC Berkeley Div Biostat Work Pap Ser*. 2012;(295):1-49.
39. VanderWeele TJ, Tchetgen Tchetgen EJ. Mediation analysis with time varying exposures and mediators. *J R Stat Soc Ser B (Statistical Methodol*. 2016;168:1-22. doi:10.1111/rssb.12194.
40. Zheng W, Van Der Laan M. Longitudinal Mediation Analysis with Time-varying Mediators and Exposures, with Application to Survival Outcomes. *J Causal Infer*. 2017. doi:10.1515/jci-2016-0006.
41. Lin S-H, Young J, Logan R, Tchetgen Tchetgen EJ, VanderWeele TJ. Parametric Mediational g-Formula Approach to Mediation Analysis with Time-varying Exposures, Mediators, and Confounders. *Epidemiology*. 2017;28(2):266-274. doi:10.1097/EDE.0000000000000609.
42. Lin S-H, Young JG, Logan R, VanderWeele TJ. Mediation analysis for a survival outcome with time-varying exposures, mediators, and confounders. *Stat Med*. 2017;(May):1-14. doi:10.1002/sim.7426.
43. Vansteelandt S, Daniel RM. Interventional Effects for Mediation Analysis with Multiple Mediators. *Epidemiology*. 2017;28(2):258-265. doi:10.1097/EDE.0000000000000596.

44. VanderWeele TJ, Robinson WR. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology*. 2014;25(4):473-484. doi:10.1097/EDE.0000000000000105.
45. Jackson JW. On the interpretation of path-specific effects in health disparities research. *Epidemiology*. 2018;(in press).
46. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15(5):615-625.
47. Pearl J. The Causal Mediation Formula-A Guide to the Assessment of Pathways and Mechanisms. *Prev Sci*. 2012;13(4):426-436. doi:10.1007/s11121-011-0270-1.
48. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychol Methods*. 2010;15(4):309-334. doi:10.1037/a0020761.